

MPI for Scalable Computing

Tutorial at ATPESC, July 2020

Latest slides and code examples are available at

<https://anl.box.com/v/atpesc2020-mpi-tutorial>

Pavan Balaji

Argonne National Laboratory

Email: balaji@anl.gov

Web: <http://www.mcs.anl.gov/~balaji>

William Gropp

University of Illinois, Urbana-Champaign

Email: wgropp@illinois.edu

Web: wgropp.cs.illinois.edu/

Yanfei Guo

Argonne National Laboratory

Email: yguo@anl.gov

Web: <https://www.mcs.anl.gov/~yguo/>

Rajeev Thakur

Argonne National Laboratory

Email: thakur@anl.gov

Web: <https://www.mcs.anl.gov/~thakur>



U.S. DEPARTMENT OF
ENERGY

About the Speakers

- **Pavan Balaji:** Computer Scientist and Group Lead, Argonne National Laboratory
 - **William Gropp:** Director, NCSA; Professor, University of Illinois, Urbana-Champaign
 - **Yanfei Guo:** Assistant Computer Scientist, Argonne National Laboratory
 - **Rajeev Thakur:** Deputy Division Director, Argonne National Laboratory
-
- All of us are deeply involved in MPI standardization (in the MPI Forum) and in MPI implementation

The MPI Part of ATPESC

- We assume everyone already has some MPI experience
- We will focus more on understanding MPI concepts than on coding details
- Emphasis will be on issues affecting scalability and performance
- There will be code walkthroughs and hands-on exercises

Outline

■ Morning

- Introduction to MPI and this tutorial
- Avoiding unnecessary synchronization
- Topics in collective communication
- Stencil example and hands-on exercises
- Minimizing data motion using MPI derived datatypes
- Begin RMA

■ Afternoon

- One-sided communication (or remote memory access)
- Hands-on exercises
- Hybrid programming
 - MPI + threads
 - MPI + shared-memory
 - MPI + GPUs
- Hands-on exercises

What is MPI?

■ MPI: Message Passing Interface

- The MPI Forum organized in 1992 with broad participation by:
 - Vendors: IBM, Intel, TMC, SGI, Convex, Meiko
 - Portability library writers: PVM, p4
 - Users: application scientists and library writers
 - MPI-1 finished in 18 months
- Incorporates the best ideas in a “standard” way
 - Each function takes fixed arguments
 - Each function has fixed semantics
 - Standardizes what the MPI implementation provides and what the application can and cannot expect
 - Each system can implement it differently as long as the semantics match

■ MPI is not...

- a language or compiler specification
- a specific implementation or product

MPI-1

- MPI-1 supports the classical message-passing programming model: basic point-to-point communication, collectives, datatypes, etc
- MPI-1 was defined (1994) by a broadly based group of parallel computer vendors, computer scientists, and applications developers.
 - 2-year intensive process
- Implementations appeared quickly and now MPI is taken for granted as vendor-supported software on any parallel machine.
- Free, portable implementations exist for clusters and other environments (MPICH, Open MPI)

Following MPI Standards

- MPI-2 was released in 1997
 - Several additional features including MPI + threads, MPI-I/O, remote memory access functionality and many others
- MPI-2.1 (2008) and MPI-2.2 (2009) were released with some corrections to the standard and small features
- MPI-3 (2012) added several new features to MPI
- MPI-3.1 (2015) is the latest version of the standard with minor corrections and features
- The Standard itself:
 - at <http://www.mpi-forum.org>
 - All MPI official releases, in both postscript and HTML
- Other information on Web:
 - at <http://www.mcs.anl.gov/mpi>
 - pointers to lots of material including tutorials, a FAQ, other MPI pages

Overview of New Features in MPI-3

- Major new features
 - Nonblocking collectives
 - Neighborhood collectives
 - Improved one-sided communication interface
 - Tools interface
 - Fortran 2008 bindings
- Other new features
 - Matching Probe and Recv for thread-safe probe and receive
 - Noncollective communicator creation function
 - “const” correct C bindings
 - Comm_split_type function
 - Nonblocking Comm_dup
 - Type_create_hindexed_block function
- C++ bindings removed
- Previously deprecated functions removed
- MPI 3.1 added nonblocking collective I/O functions

Status of MPI-3.1 Implementations

	MPICH	MPVAPICH	Open MPI	Cray	Tianhe	Intel		IBM			HPE	Fujitsu	MS	MPC	NEC	Sunway	RIKEN	AMPI
						IMPI	MPICH-OFI	BG/Q (legacy) ¹	PE (legacy) ²	Spectrum								
NBC	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Nbr. Coll.	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓	✓	✓
RMA	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	(*)	✓	✓	✓	✓	Q2 '18
Shr. mem	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	Q1 '18
MPI_T	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	*	✓	✓	✓	✓	Q2 '18
Comm-create group	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	*	✓	✓	✓	✓	✓
F08 Bindings	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✗	✗	✓	✓	✓	✓	Q2 '18
New Dtypes	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Large Counts	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
MProbe	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	Q1 '18
NBC I/O	✓	✓	✓	✓	✗	✓	✓	✗	✗	✓	✓	✗	✗	*	✓	✗	✓	Q3 '18

Release dates are estimates; subject to change at any time

“✗” indicates no publicly announced plan to implement/support that feature

Platform-specific restrictions might apply to the supported features

¹ Open Source but unsupported

² No MPI_T variables exposed

* Under development

(*) Partly done

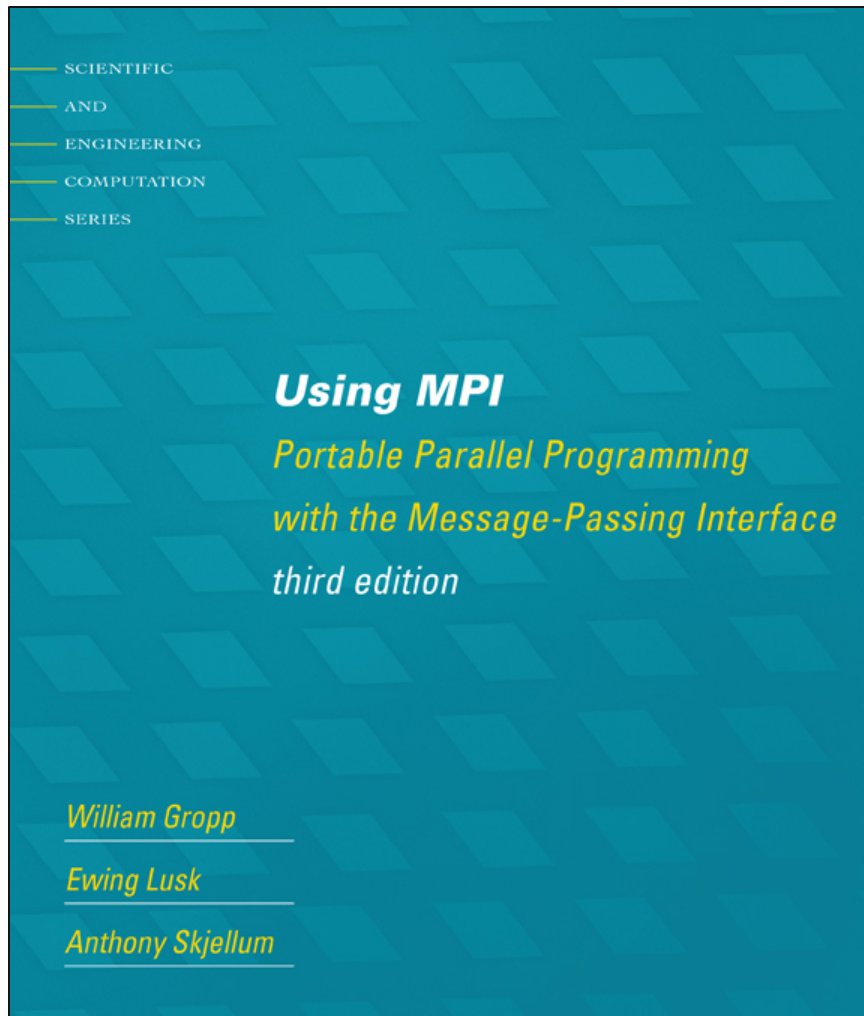
Important considerations while using MPI

- All parallelism is explicit: the programmer is responsible for correctly identifying parallelism and implementing parallel algorithms using MPI constructs

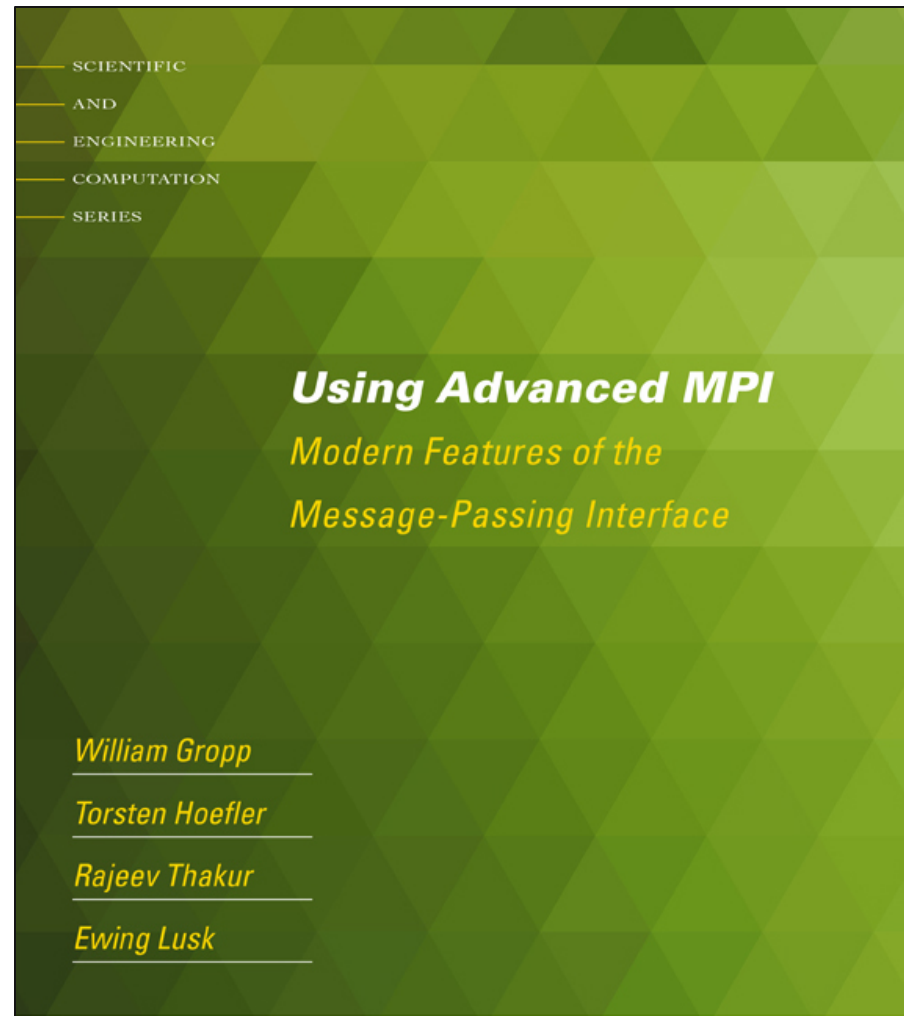
Web Pointers

- MPI standard : <http://www.mpi-forum.org/docs/docs.html>
- MPI Forum : <http://www.mpi-forum.org/>
- MPI implementations:
 - MPICH : <http://www.mpich.org>
 - MVAPICH : <http://mvapich.cse.ohio-state.edu/>
 - Intel MPI: <http://software.intel.com/en-us/intel-mpi-library/>
 - Microsoft MPI: <https://docs.microsoft.com/en-us/message-passing-interface/microsoft-mpi>
 - Open MPI : <http://www.open-mpi.org/>
 - IBM MPI, Cray MPI, HP MPI, TH MPI, ...
- Several MPI tutorials can be found on the web

Tutorial Books on MPI



Basic MPI

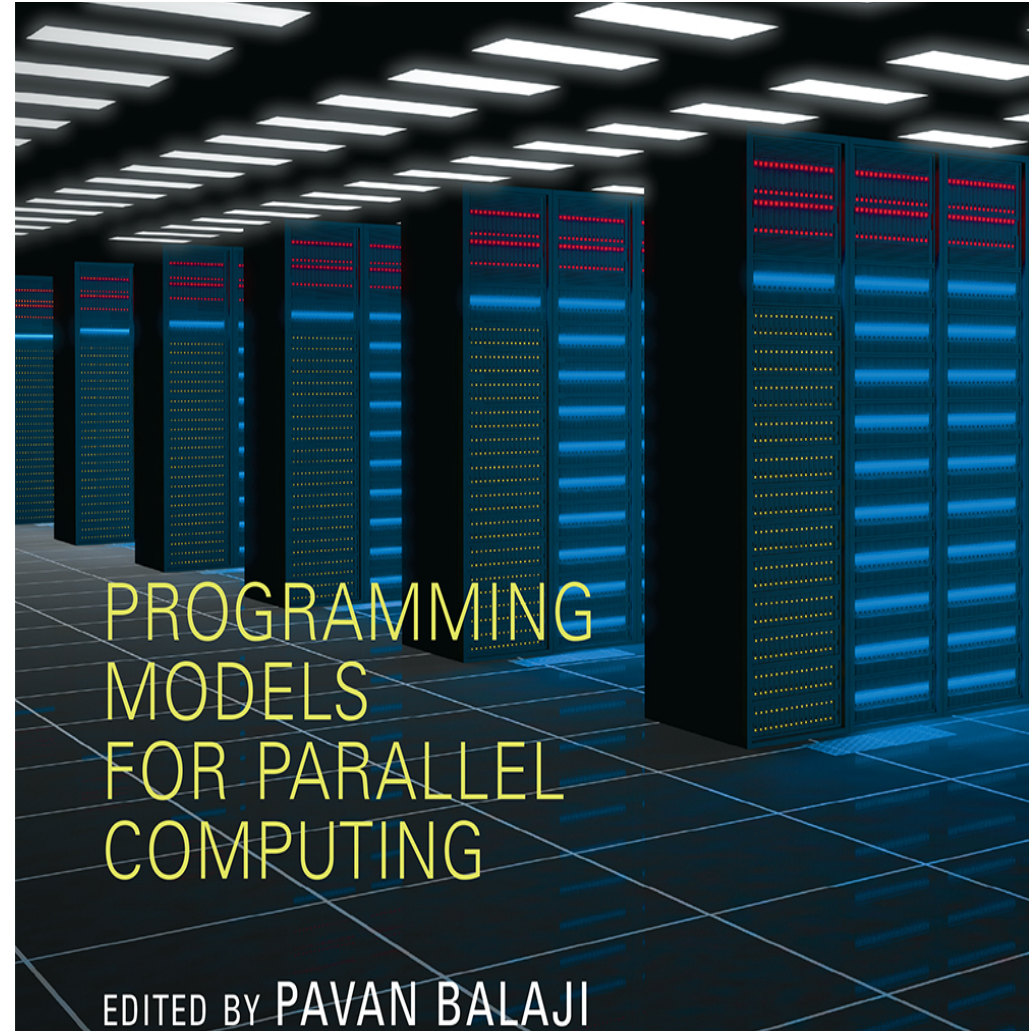


Advanced MPI, including MPI-3

Book on Parallel Programming Models

Edited by Pavan Balaji

- **MPI:** W. Gropp and R. Thakur
- **GASNet:** P. Hargrove
- **OpenSHMEM:** J. Kuehn and S. Poole
- **UPC:** K. Yelick and Y. Zheng
- **Global Arrays:** S. Krishnamoorthy, J. Daily, A. Vishnu, and B. Palmer
- **Chapel:** B. Chamberlain
- **Charm++:** L. Kale, N. Jain, and J. Lifflander
- **ADLB:** E. Lusk, R. Butler, and S. Pieper
- **Scioto:** J. Dinan
- **SWIFT:** T. Armstrong, J. M. Wozniak, M. Wilde, and I. Foster
- **CnC:** K. Knobe, M. Burke, and F. Schlimbach
- **OpenMP:** B. Chapman, D. Eachempati, and S. Chandrasekaran
- **Cilk Plus:** A. Robison and C. Leiserson
- **Intel TBB:** A. Kukanov
- **CUDA:** W. Hwu and D. Kirk
- **OpenCL:** T. Mattson



Approach in this Tutorial

- Example driven
 - A few running examples used throughout the tutorial
 - Other smaller examples used to illustrate specific features

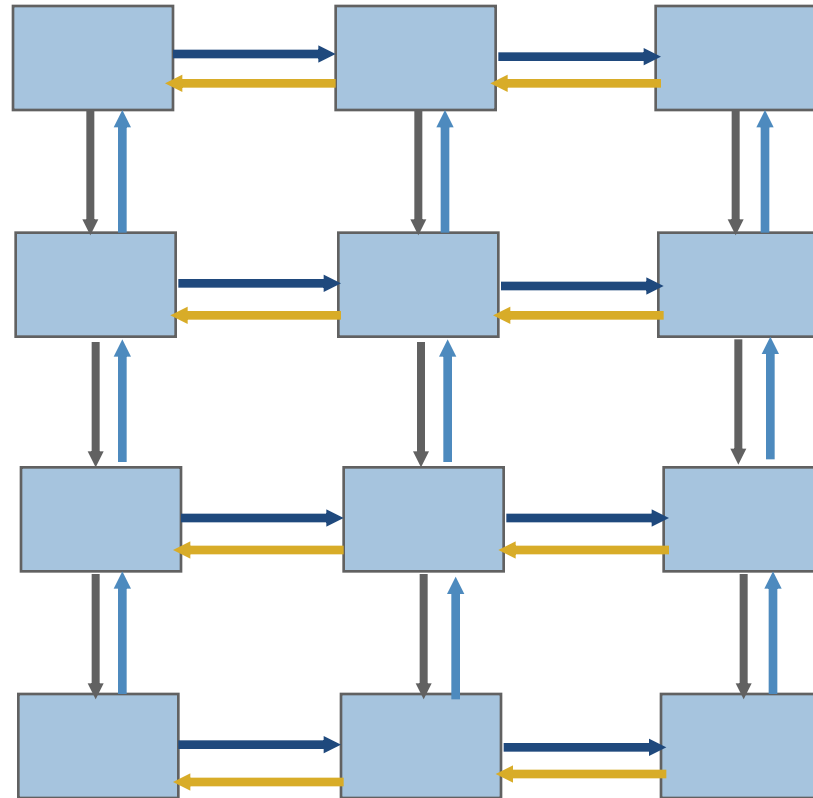
Costs of Unintended Synchronization

Unexpected Hot Spots

- Even simple operations can give surprising performance behavior
- Examples arise even in common grid exchange patterns
- Message passing illustrates problems present even in shared memory
 - Blocking operations may cause unavoidable stalls

Mesh Exchange

- Exchange data on a mesh



Sample Code

- Do i=1,n_neighbors
 Call MPI_Send(edge(1,i), len, MPI_REAL, &nbr(i), tag, comm, ierr)
Enddo

Do i=1,n_neighbors
 Call MPI_Recv(edge(1,i), len, MPI_REAL, &nbr(i), tag, comm, status, ierr)
Enddo

Deadlocks!

- All of the sends may block, waiting for a matching receive (will for large enough messages)
- The variation of
if (has down nbr) then
 Call MPI_Send(... down ...)
endif
if (has up nbr) then
 Call MPI_Recv(... up ...)
endif
...
sequentializes (all except the bottom process blocks)

Sequentialization

Start Send	Start Send	Start Send	Start Send	Start Send	Start Send Send	Send Recv	Recv
				Send	Recv		
			Send	Recv			
	Send	Recv					
Send	Recv						

Fix 1: Use Irecv

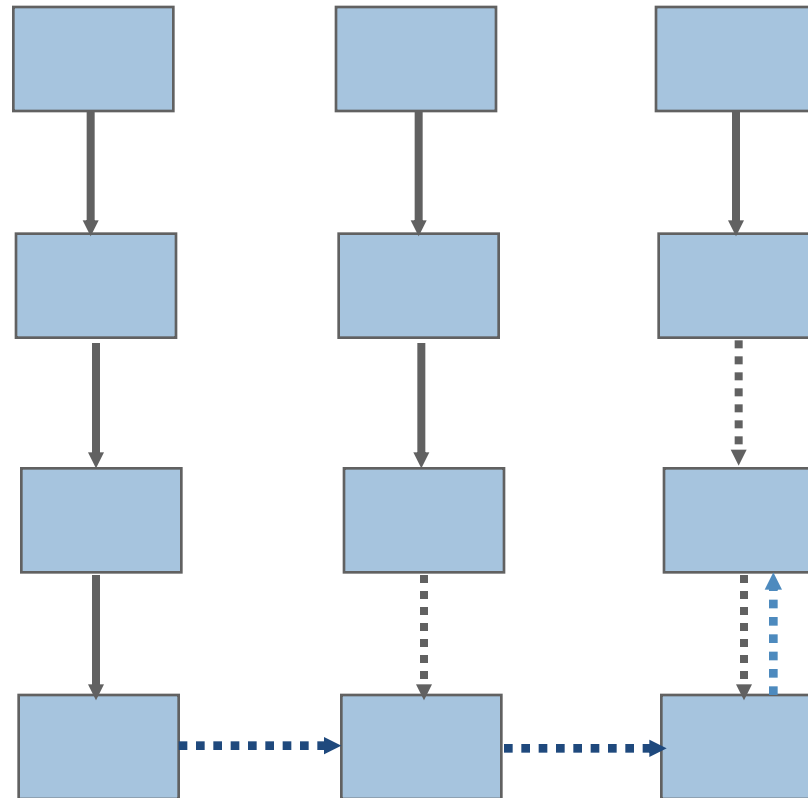
- Do i=1, n_neighbors
 Call MPI_Irecv(inedge(1,i), len, MPI_REAL, nbr(i), tag, &comm, requests(i), ierr)
Enddo
Do i=1, n_neighbors
 Call MPI_Send(edge(1,i), len, MPI_REAL, nbr(i), tag, &comm, ierr)
Enddo
Call MPI_Waitall(n_neighbors, requests, statuses, ierr)
- Does not perform well in practice. Why?

Understanding the Behavior: Timing Model

- Sends interleave
- Sends block (data larger than buffering will allow)
- Sends control timing
- Receives do not interfere with Sends
- Exchange can be done in 4 steps (down, right, up, left)

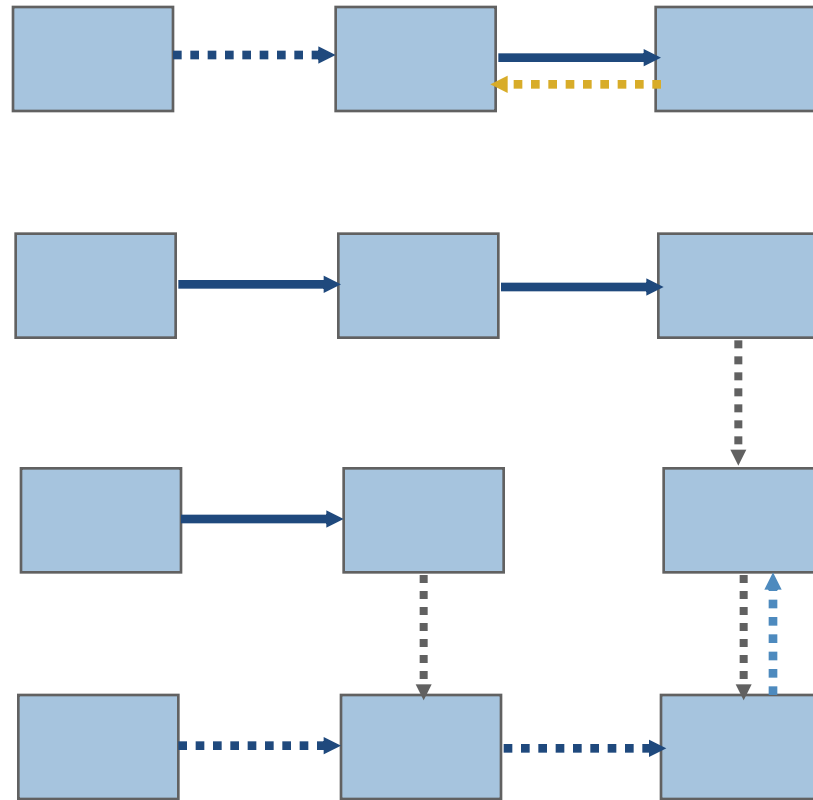
Mesh Exchange - Step 1

- Exchange data on a mesh



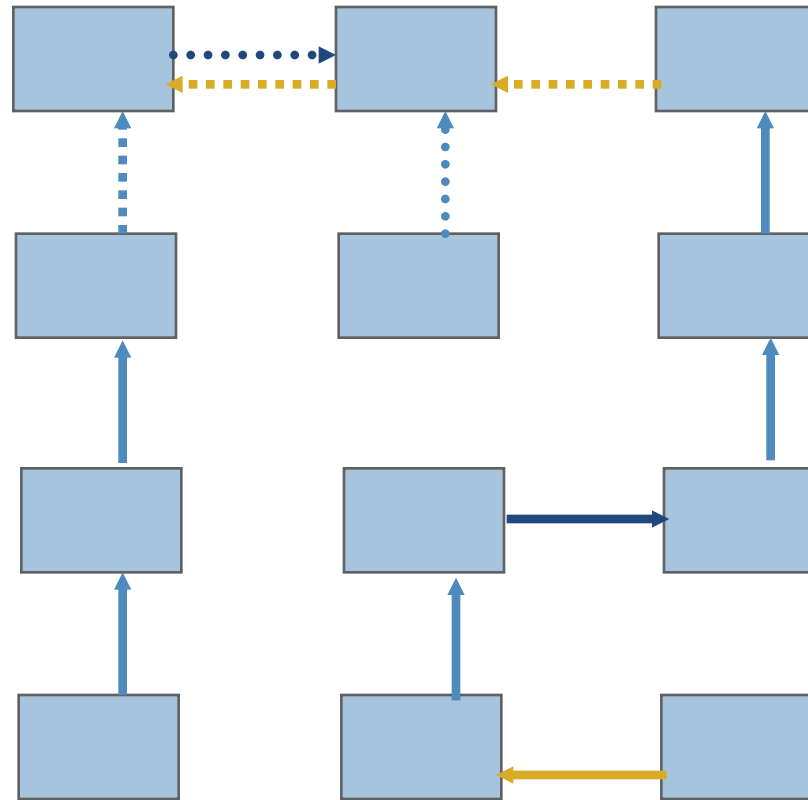
Mesh Exchange - Step 2

- Exchange data on a mesh



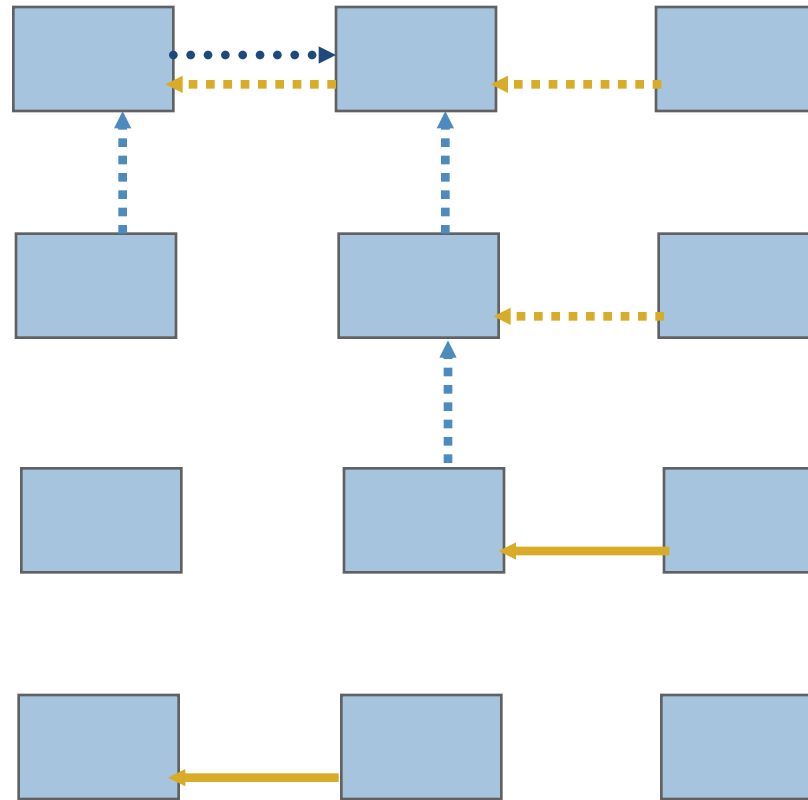
Mesh Exchange - Step 3

- Exchange data on a mesh



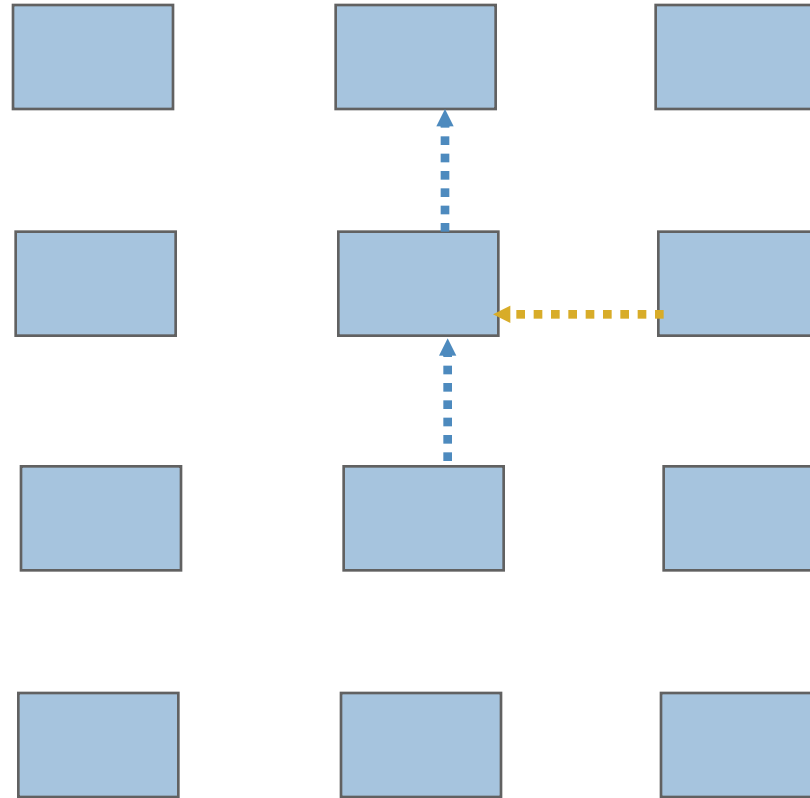
Mesh Exchange - Step 4

- Exchange data on a mesh



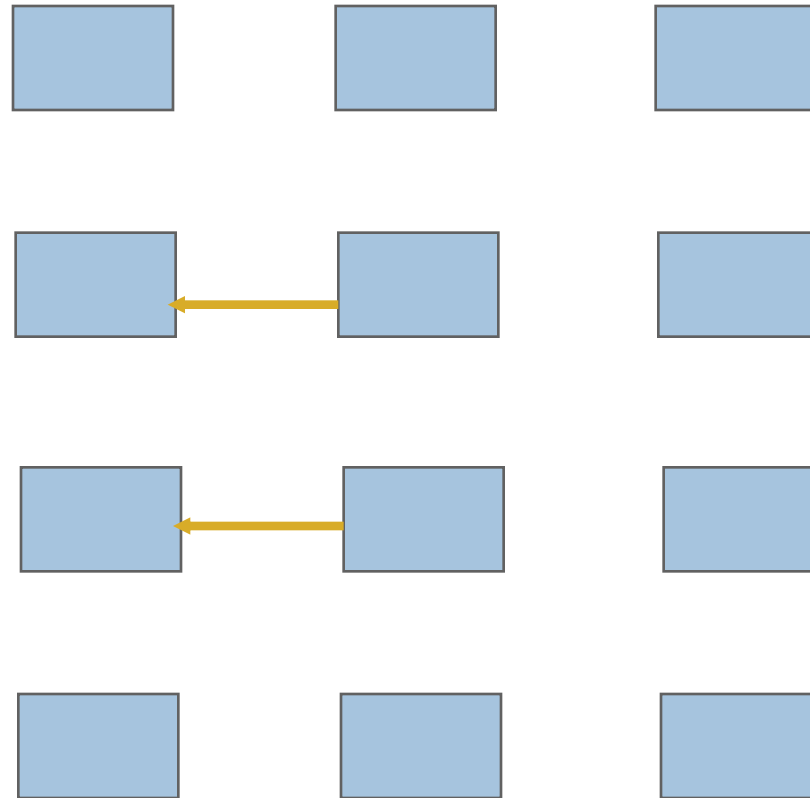
Mesh Exchange - Step 5

- Exchange data on a mesh

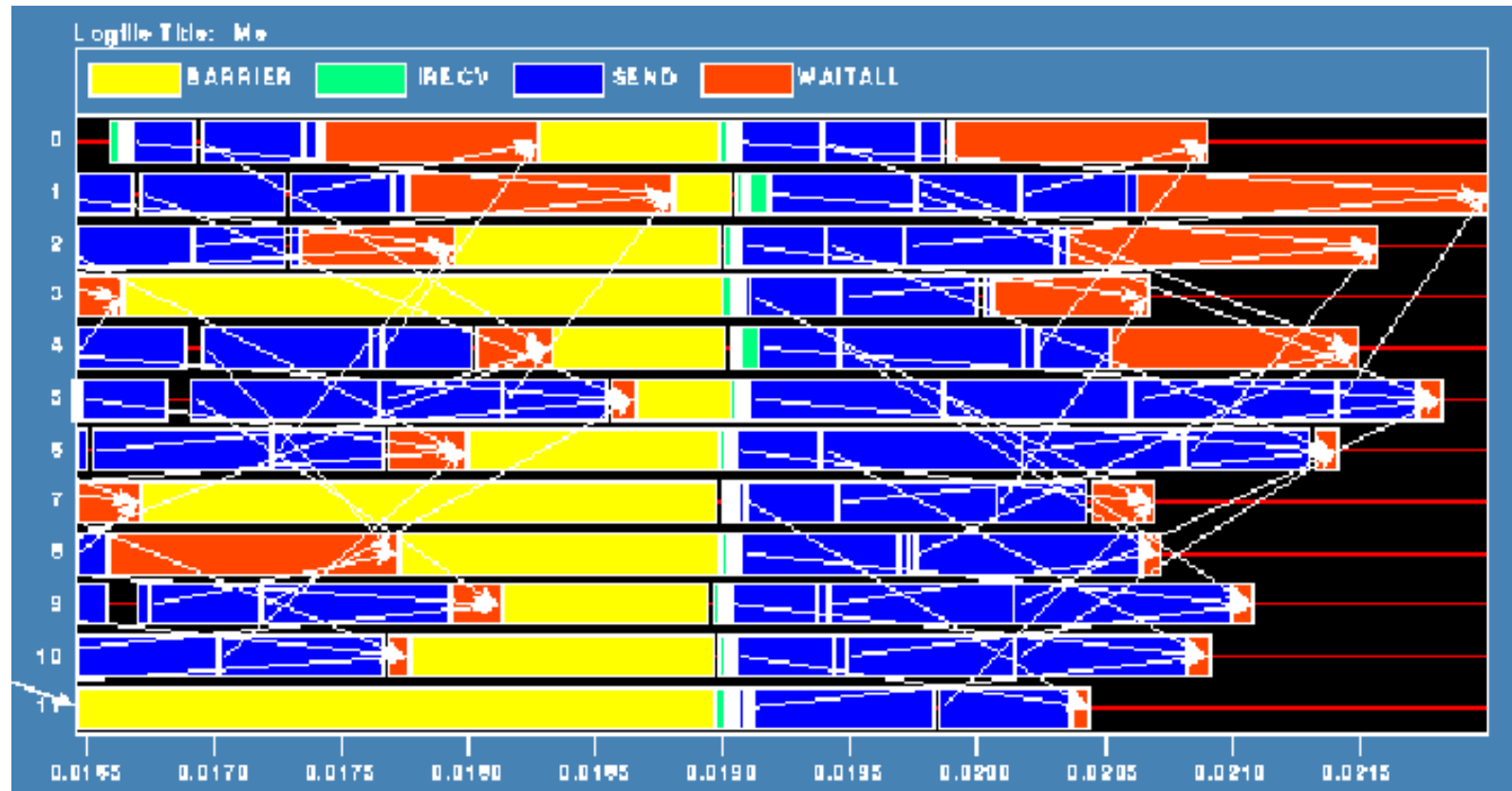


Mesh Exchange - Step 6

- Exchange data on a mesh

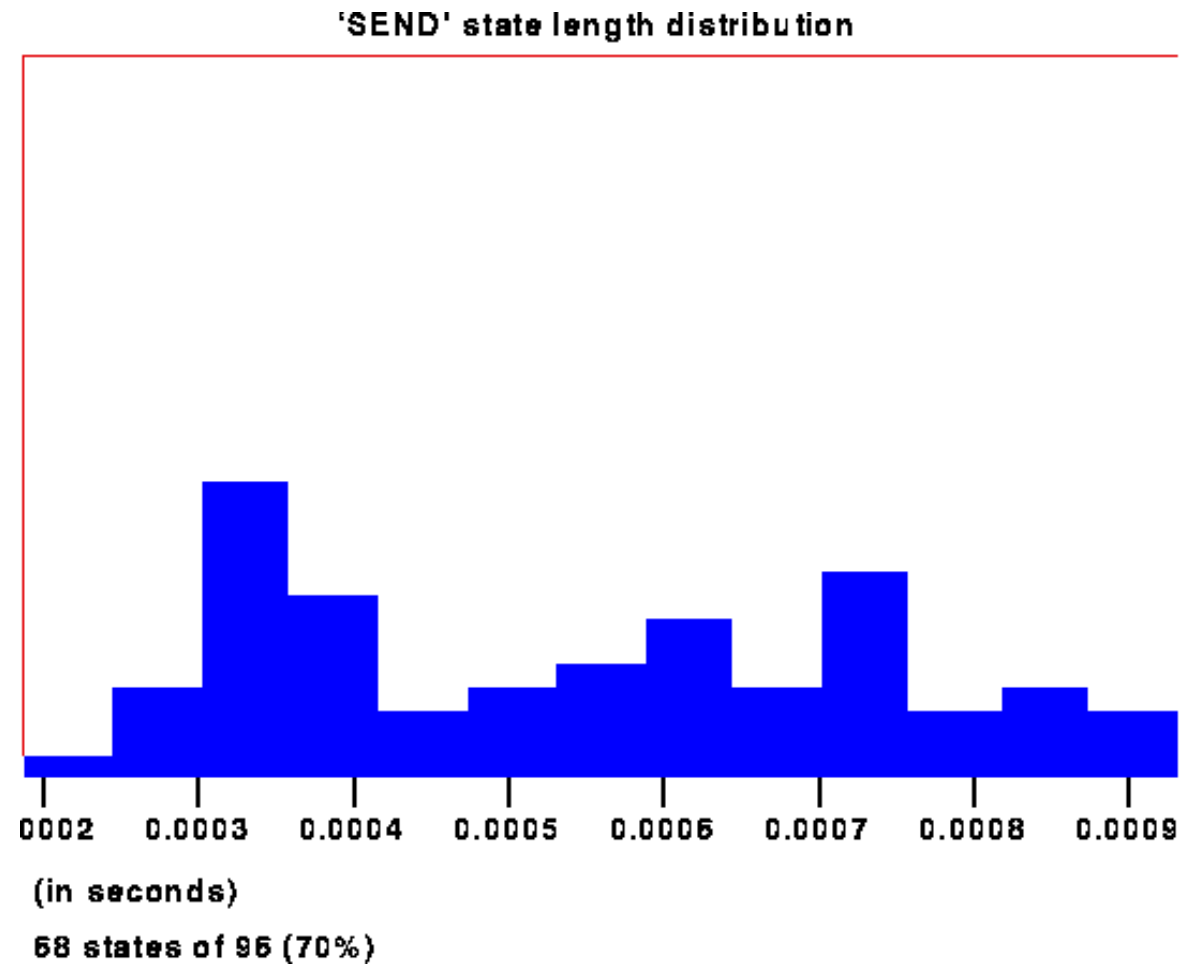


Timeline



- Note that process 1 finishes last, as predicted

Distribution of Sends



Why Six Steps?

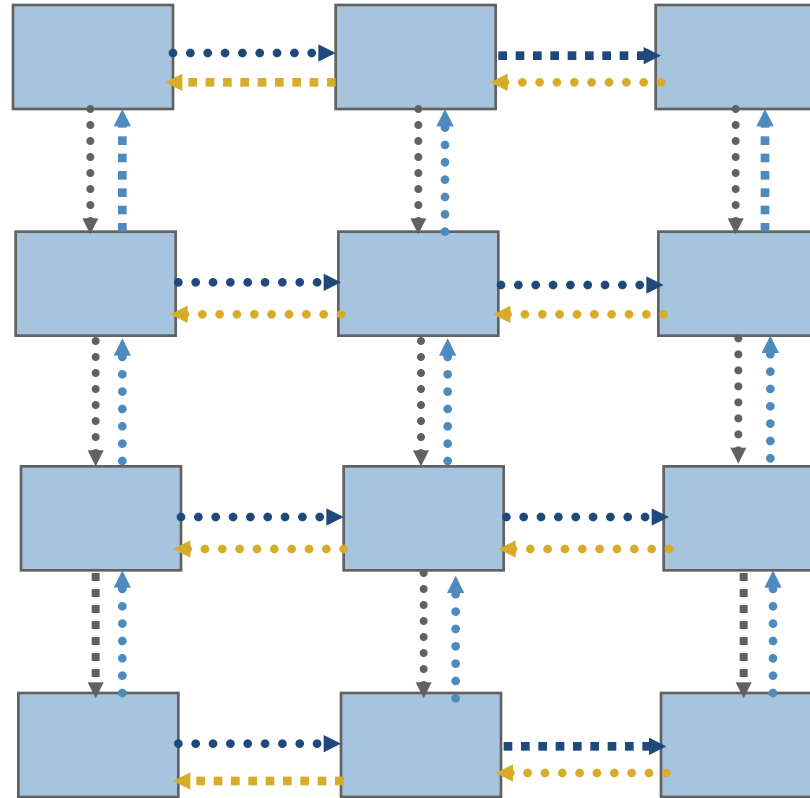
- Ordering of Sends introduces delays when there is contention at the receiver
- Takes roughly twice as long as it should
- Bandwidth is being wasted
- Same thing would happen if using memcpy and shared memory

Fix 2: Use Isend and Irecv

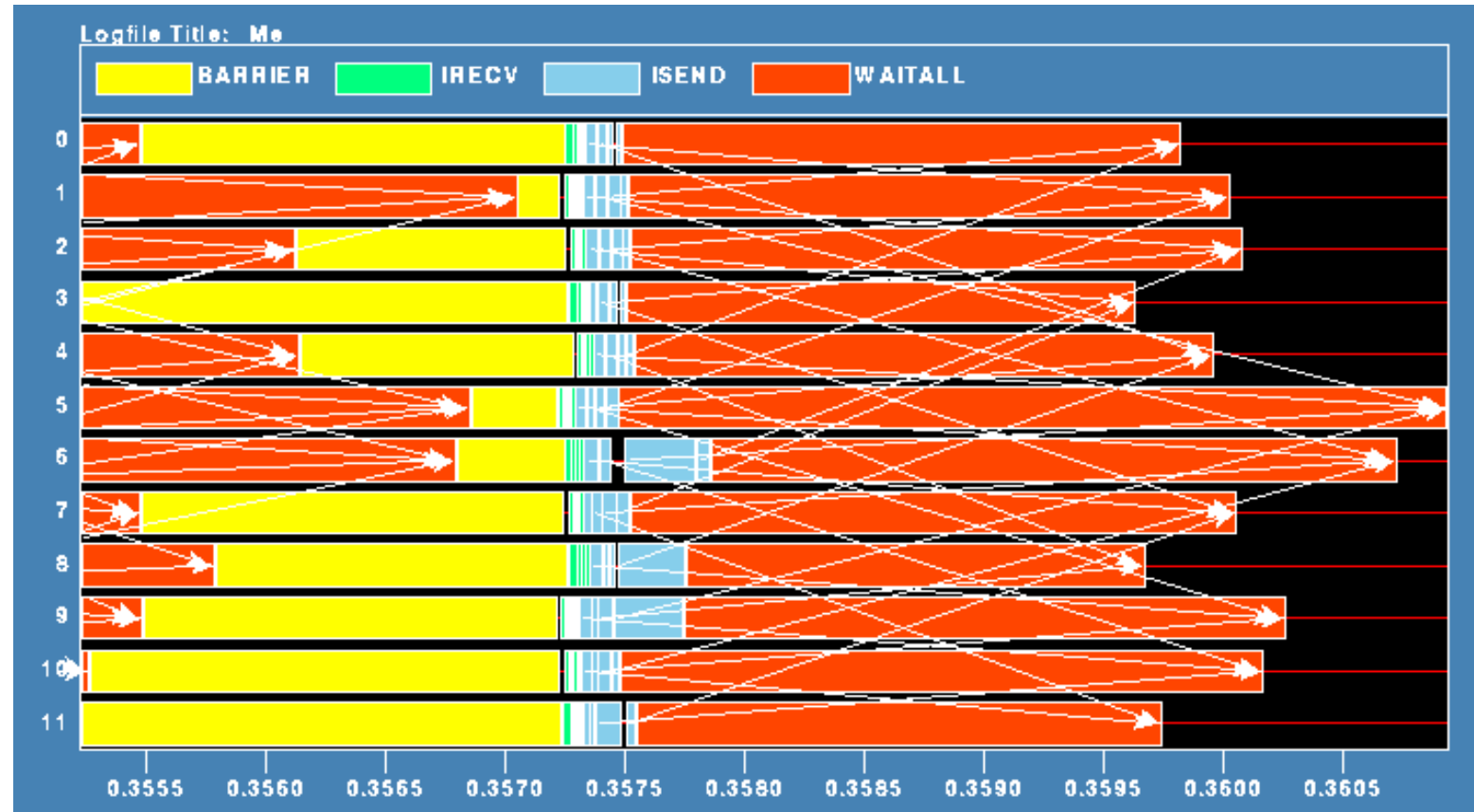
- Do i=1, n_neighbors
 Call MPI_Irecv(inedge(1,i), len, MPI_REAL, nbr(i), tag, &comm, requests(i), ierr)
Enddo
Do i=1,n_neighbors
 Call MPI_Isend(edge(1,i), len, MPI_REAL, nbr(i), tag, &comm,
 requests(n_neighbors+i), ierr)
Enddo
Call MPI_Waitall(2*n_neighbors, requests, statuses, ierr)

Mesh Exchange - Steps 1-4

- Four interleaved steps



Timeline with Isend-Irecv



Note processes 5 and 6 are the only interior processes; these perform more communication than the other processes

Lesson: Defer Synchronization

- Send-receive accomplishes two things:
 - Data transfer
 - Synchronization
- In many cases, there is more synchronization than required
- Consider the use of nonblocking operations and `MPI_Waitall` to defer synchronization
 - Effectiveness depends on how data is moved by the MPI implementation
 - E.g., If large messages are moved by blocking RMA operations “under the covers,” the implementation can’t adapt to contention at the target processes, and you may see no benefit.
 - This is more likely with larger messages

Collectives: Blocking and Nonblocking



Introduction to Collective Operations in MPI

- Collective operations are called by all processes in a communicator.
- **MPI_BCAST** distributes data from one process (the root) to all others in a communicator.
- **MPI_REDUCE** combines data from all processes in the communicator and returns it to one process.
- In many numerical algorithms, **SEND/RECV** can be replaced by **BCAST/REDUCE**, improving both simplicity and efficiency.

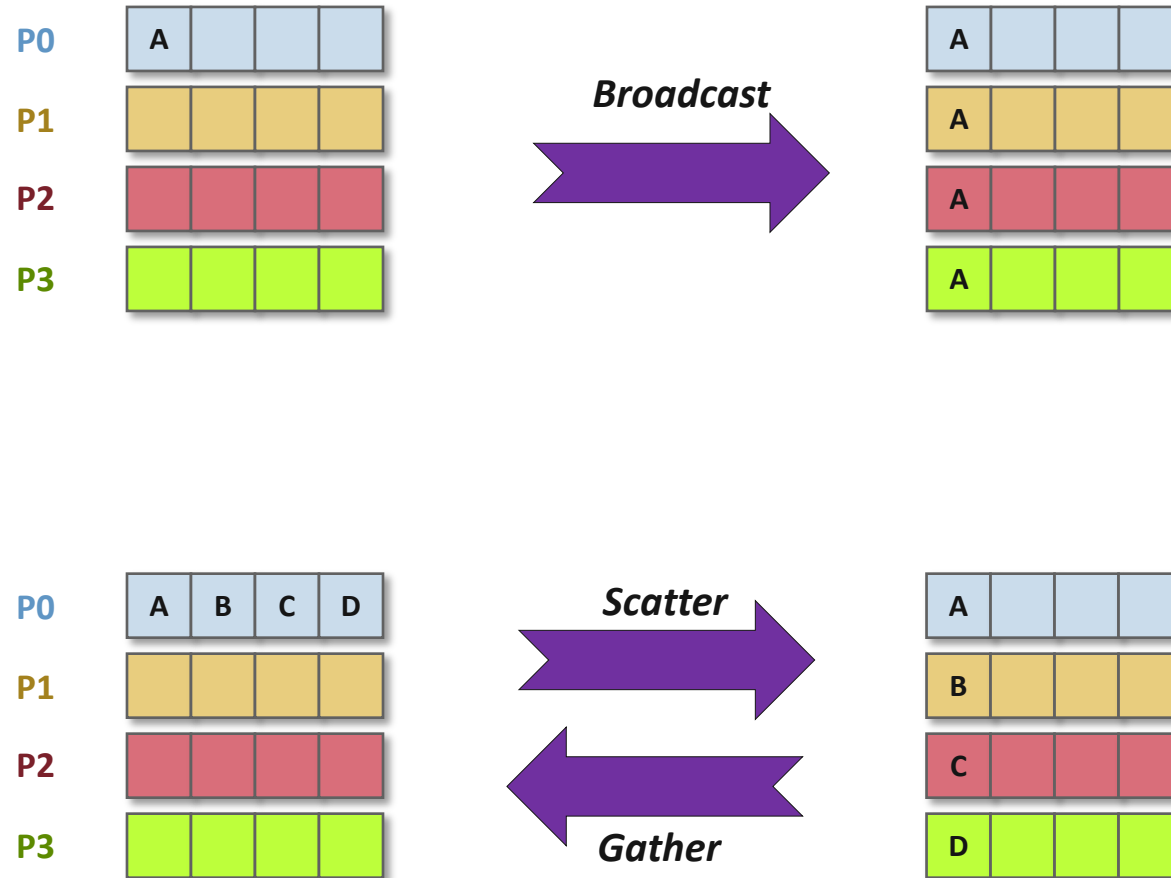
MPI Collective Communication

- Communication and computation is coordinated among a group of processes in a communicator
- Tags are not used; different communicators deliver similar functionality
- Non-blocking collective operations in MPI-3
- Three classes of operations: synchronization, data movement, collective computation

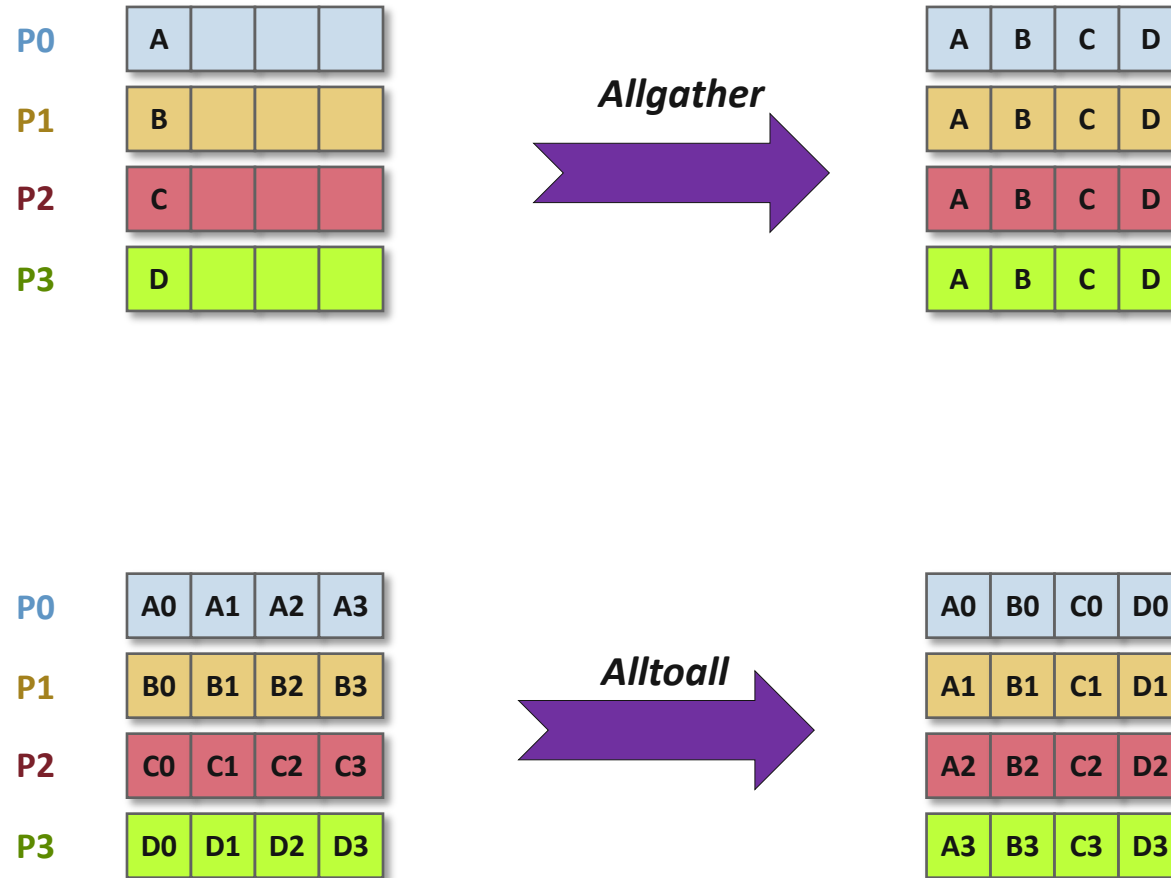
Synchronization

- **MPI_BARRIER(comm)**
 - Blocks until all processes in the group of the communicator **comm** call it
 - A process cannot get out of the barrier until all other processes have reached barrier

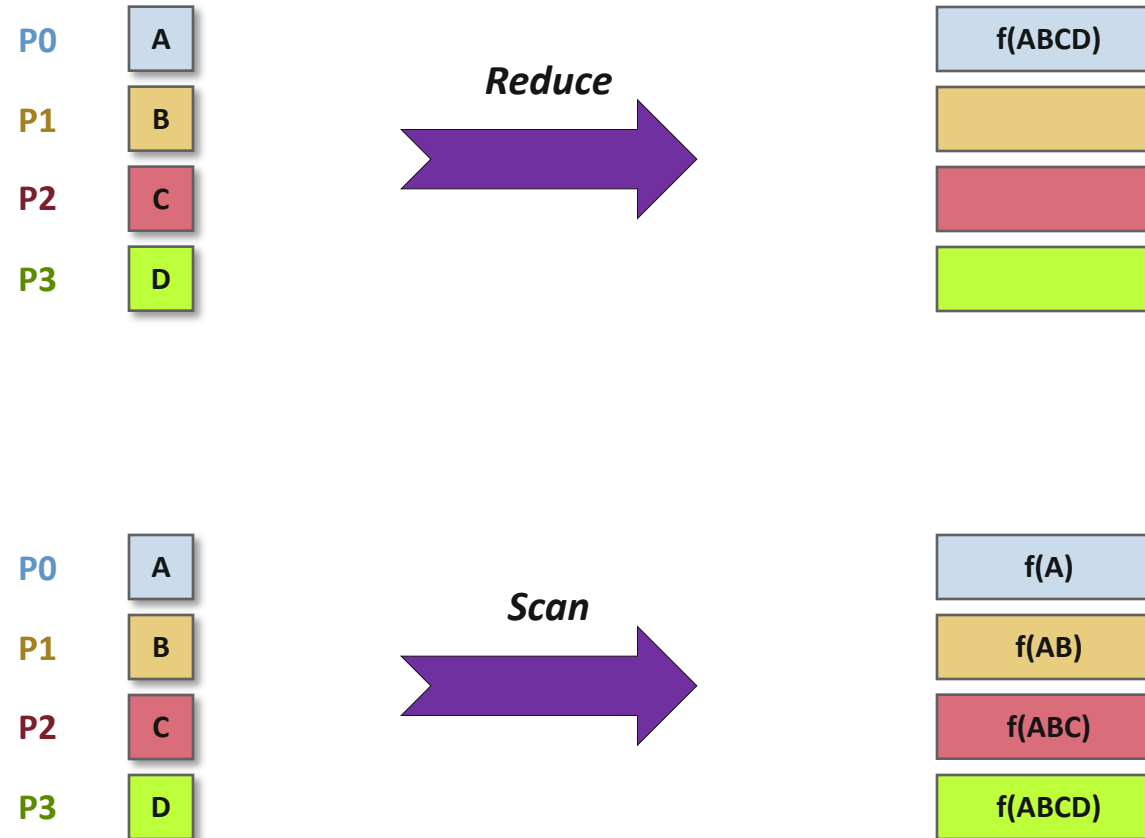
Collective Data Movement



More Collective Data Movement



Collective Computation



MPI Collective Routines

- Many Routines: MPI_**ALL**GATHER, MPI_**ALL**GATHER**V**, MPI_**ALL**REDUCE, MPI_**ALL**TOALL, MPI_**ALL**TOALL**V**, MPI_BCAST, MPI_GATHER, MPI_GATHER**V**, MPI_REDUCE, MPI_REDUCE**SCATTER**, MPI_SCAN, MPI_SCATTER, MPI_SCATTER**V**
- “**A**ll” versions deliver results to all participating processes
- “**V**” versions (stands for vector) allow the chunks to have different sizes
- MPI_**ALL**REDUCE, MPI_REDUCE, MPI_REDUCE**SCATTER**, and MPI_SCAN take both built-in and user-defined combiner functions

MPI Built-in Collective Computation Operations

▪ <code>MPI_MAX</code>	Maximum
▪ <code>MPI_MIN</code>	Minimum
▪ <code>MPI_PROD</code>	Product
▪ <code>MPI_SUM</code>	Sum
▪ <code>MPI_LAND</code>	Logical and
▪ <code>MPI_LOR</code>	Logical or
▪ <code>MPI_LXOR</code>	Logical exclusive or
▪ <code>MPI_BAND</code>	Bitwise and
▪ <code>MPI_BOR</code>	Bitwise or
▪ <code>MPI_BXOR</code>	Bitwise exclusive or
▪ <code>MPI_MAXLOC</code>	Maximum and location
▪ <code>MPI_MINLOC</code>	Minimum and location

Defining your own Collective Operations

- Create your own collective computations with:

```
MPI_OP_CREATE(user_fn, commutes, &op);  
MPI_OP_FREE(&op);  
  
user_fn(invec, inoutvec, len, datatype);
```

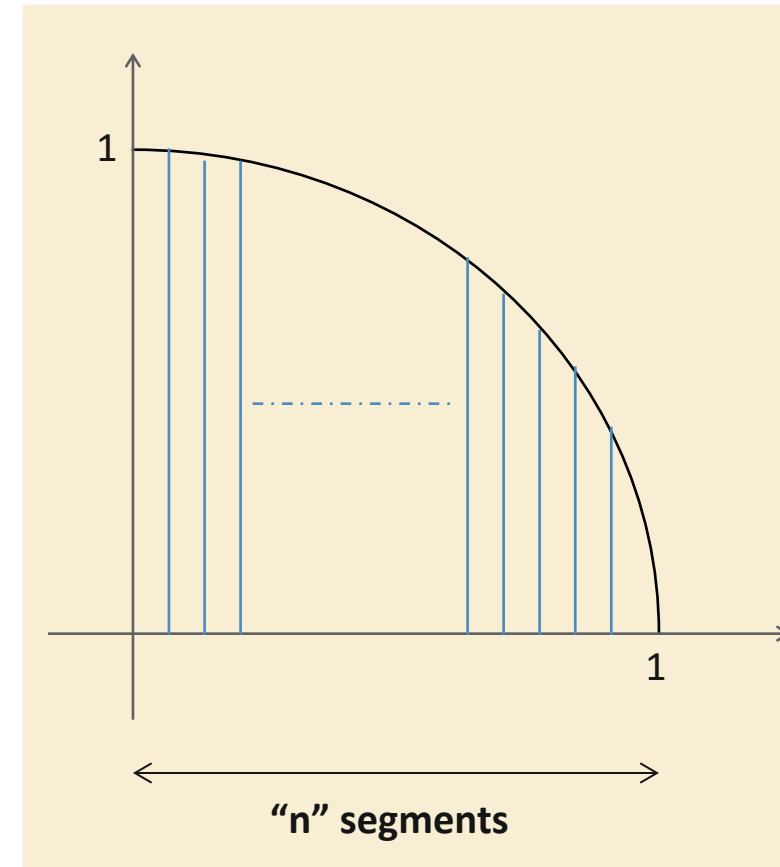
- The user function should perform:

```
for i from 0 to len-1  
    inoutvec[i] = invec[i] op inoutvec[i];
```

- The user function can be non-commutative, but must be associative

Example: Calculating Pi (1/3)

- Calculating the value of “pi” via numerical integration
 - Divide interval up into subintervals
 - Assign subintervals to processes
 - Each process calculates partial sum
 - Add all the partial sums together to get pi



1. Width of each segment (w) will be $1/n$
2. Distance ($d(i)$) of segment “ i ” from the origin will be “ $i * w$ ”
3. Height of segment “ i ” will be $\sqrt{1 - [d(i)]^2}$

Example: Calculating Pi (2/3)

```
#include <mpi.h>
#include <math.h>
int main(int argc, char *argv[])
{
    [...snip...]
    /* Tell all processes, the number of segments you want */
    MPI_Bcast(&n, 1, MPI_INT, 0, MPI_COMM_WORLD);
    w = 1.0 / (double) n;
    mypi = 0.0;
    for (i = rank + 1; i <= n; i += size)
        mypi += w * sqrt(1 - (((double) i / n) * ((double) i / n)));
    MPI_Reduce(&mypi, &pi, 1, MPI_DOUBLE, MPI_SUM, 0,
               MPI_COMM_WORLD);
    if (rank == 0)
        printf("pi is approximately %.16f, Error is %.16f\n", 4 * pi,
               fabs((4 * pi) - PI25DT));
    [...snip...]
}
```

Example: Calculating Pi (3/3)

- *blocking_coll/cpi.c*
- Calculating the approximate value of PI in parallel

Nonblocking Collective Communication

- Nonblocking (send/recv) communication
 - Deadlock avoidance
 - Overlapping communication/computation
- Collective communication
 - Collection of pre-defined optimized routines
- → Nonblocking collective communication
 - Combines both techniques
 - System noise/imbalance resiliency
 - Semantic advantages

Nonblocking Collective Communication

- Nonblocking variants of all collectives
 - `MPI_Ibcast(<bcast args>, MPI_Request *req);`
- Semantics
 - Function returns no matter what
 - No guaranteed progress (quality of implementation)
 - Usual completion calls (wait, test) + mixing
 - Out-of order completion
- Restrictions
 - Send and vector buffers may not be updated during operation (like other nonblocking operations)
 - No tags, in-order matching (like other collective operations)
 - `MPI_Cancel` not supported
 - No matching with blocking collectives

Nonblocking Collective Communication

- Semantic advantages
 - Enable asynchronous progression (and manual)
 - Software pipelining
 - Decouple data transfer and synchronization
 - Noise resiliency!
 - Allow overlapping communicators
 - See also neighborhood collectives
 - Multiple outstanding operations at any time
 - Enables pipelining window

A Nonblocking Barrier?

- Semantics:
 - MPI_Ibarrier() – calling process entered the barrier, **no** synchronization happens
 - Synchronization **may** happen asynchronously
 - MPI_Test/Wait() – synchronization happens **if** necessary
- Uses:
 - Overlap barrier latency (small benefit)
 - Use the split semantics! Processes **notify** noncollectively but **synchronize** collectively!

Section Summary

- Collectives are a very powerful feature in MPI
- Optimized heavily in most MPI implementations
 - Algorithmic optimizations (e.g., tree-based communication)
 - Hardware optimizations (e.g., network or switch-based collectives)
- Matches the communication pattern of many applications
- Nonblocking collectives combine the semantics of nonblocking point-to-point and blocking collectives
- Natural extension to blocking collectives for event-driven programming
- Hardware implementations already exist for most (but not all) nonblocking collectives

Running Example: Stencil

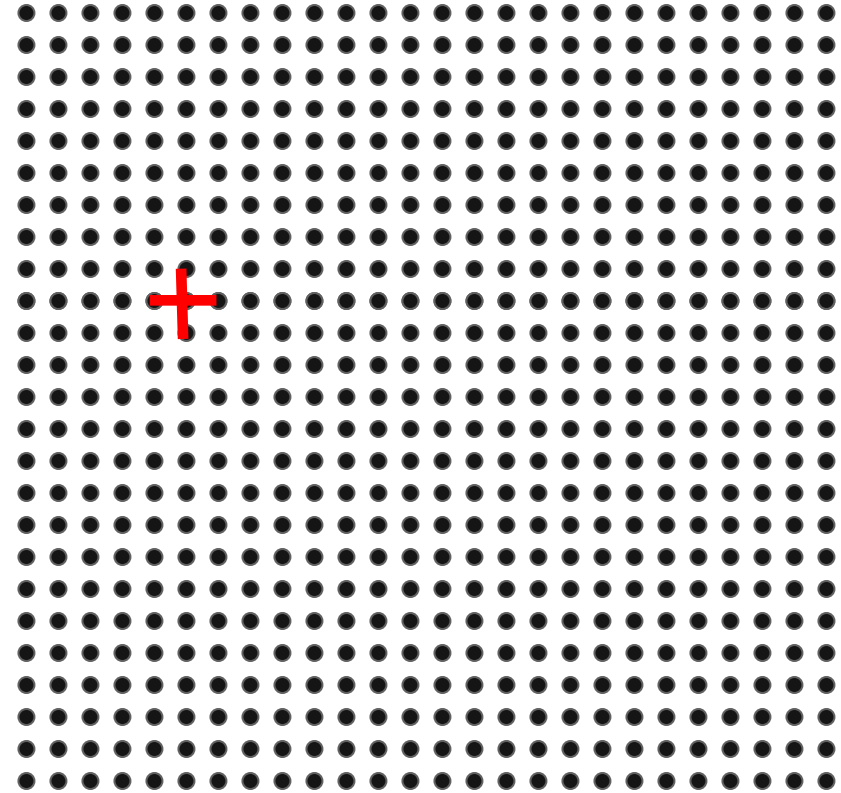


Running Example: Regular Mesh Algorithms

- Many scientific applications involve the solution of partial differential equations (PDEs)
- Many algorithms for approximating the solution of PDEs rely on forming a set of difference equations
 - Finite difference, finite elements, finite volume
- The exact form of the differential equations depends on the particular method
 - From the point of view of parallel programming for these algorithms, the operations are the same
- Five-point stencil is a popular approximation solution

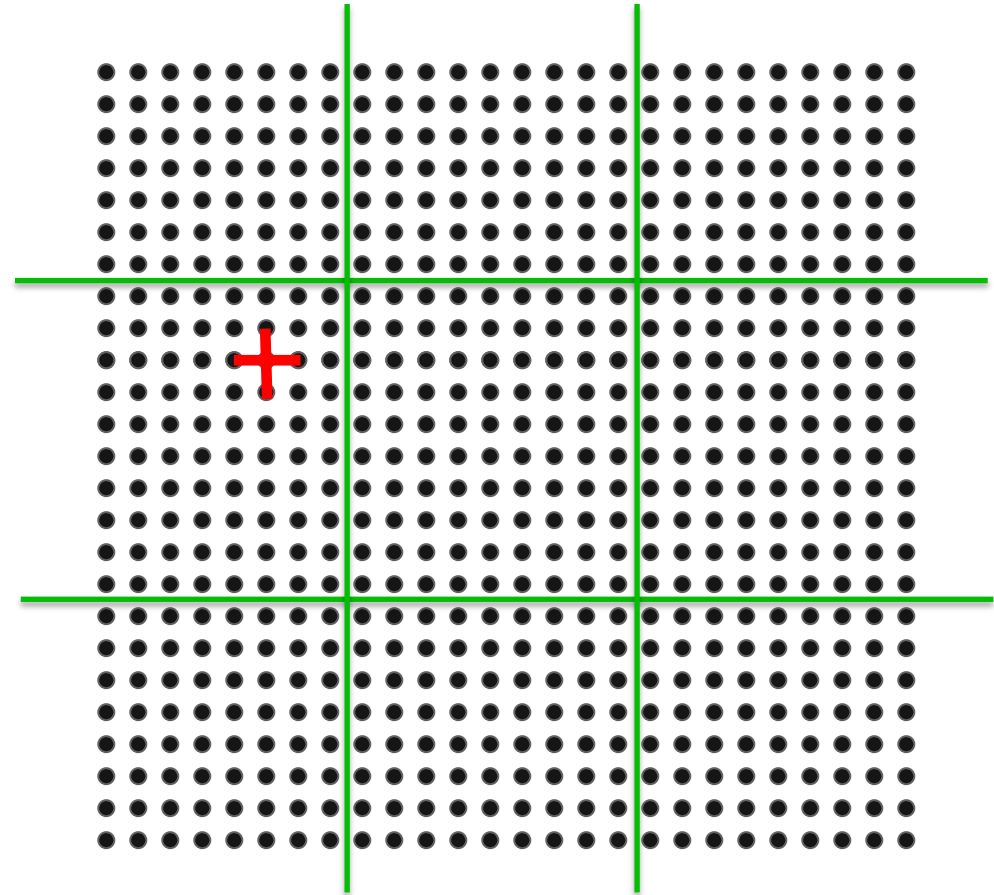
The Global Data Structure

- Each circle is a mesh point
- Difference equation evaluated at each point involves the four neighbors
- The red “plus” is called the method’s stencil
- Good numerical algorithms form a matrix equation $Au=f$; solving this requires computing Bv , where B is a matrix derived from A . These evaluations involve computations with the neighbors on the mesh.



The Global Data Structure

- Each circle is a mesh point
- Difference equation evaluated at each point involves the four neighbors
- The red “plus” is called the method’s stencil
- Good numerical algorithms form a matrix equation $Au=f$; solving this requires computing Bv , where B is a matrix derived from A . These evaluations involve computations with the neighbors on the mesh.
- Decompose mesh into equal sized (work) pieces



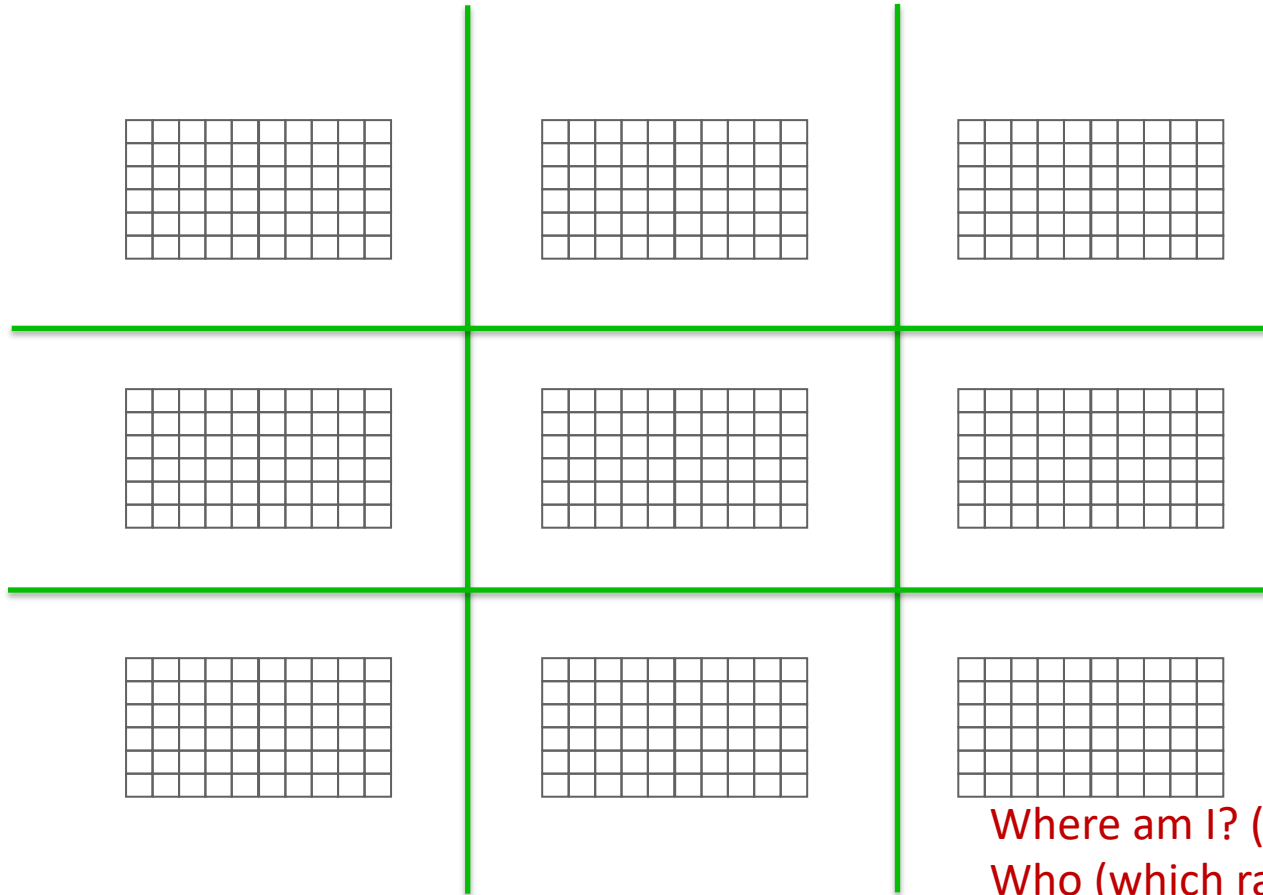
Domain Decompositioin

Parameters for domain decomposition:

N = Size of the edge of the global problem domain (assuming square)

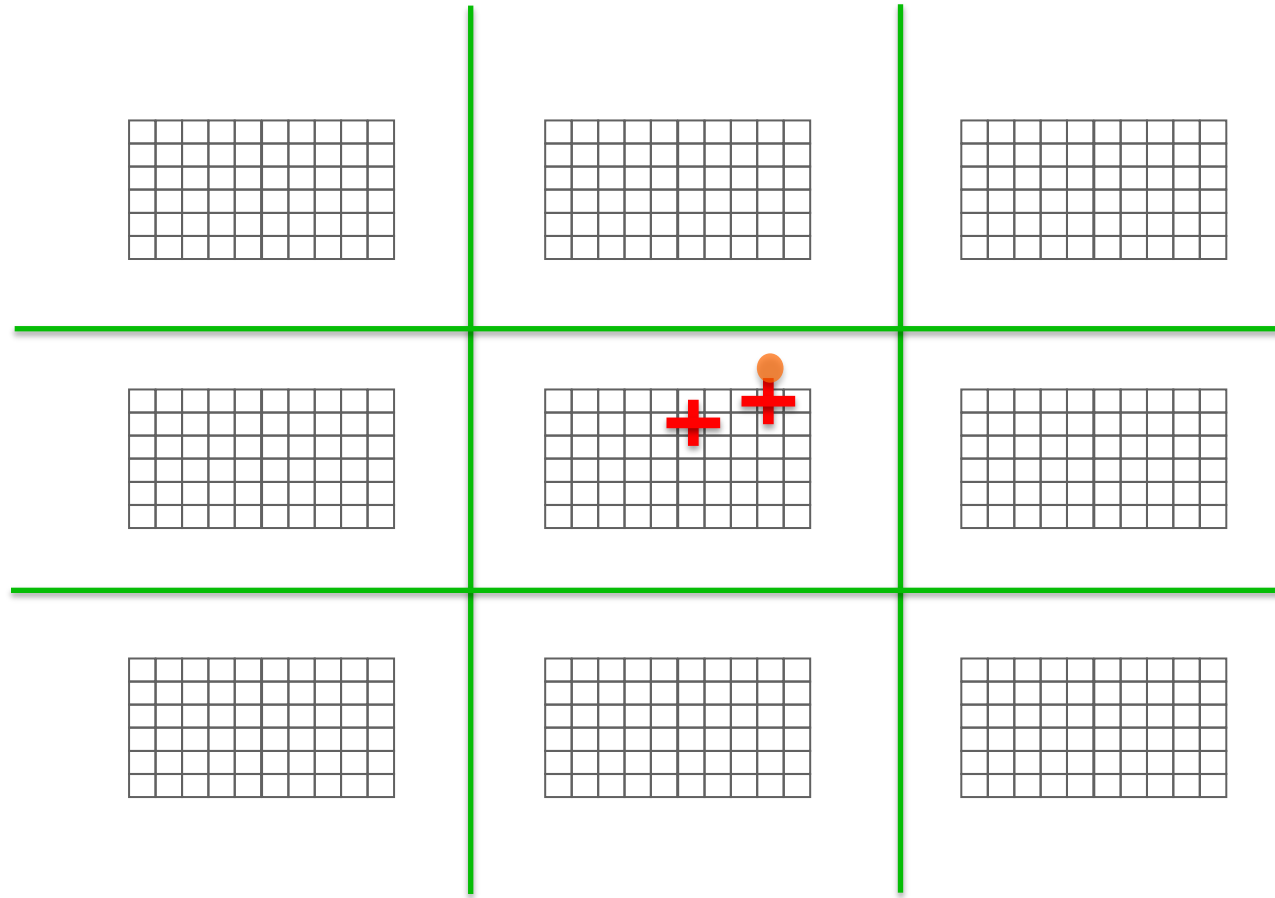
PX, PY = Number of processes in X and Y dimension

$N \% PX == 0, N \% PY == 0$



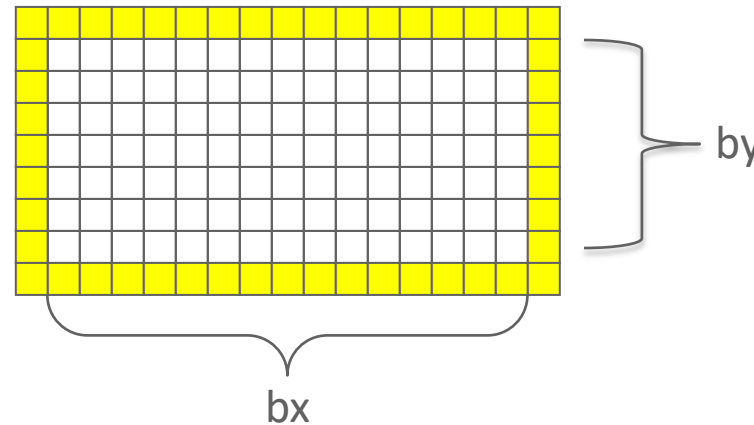
Where am I? (Global offset)
Who (which ranks) are my neighbors?
Use `MPI_PROC_NULL` for boundary

Necessary Data Transfers



The Local Data Structure

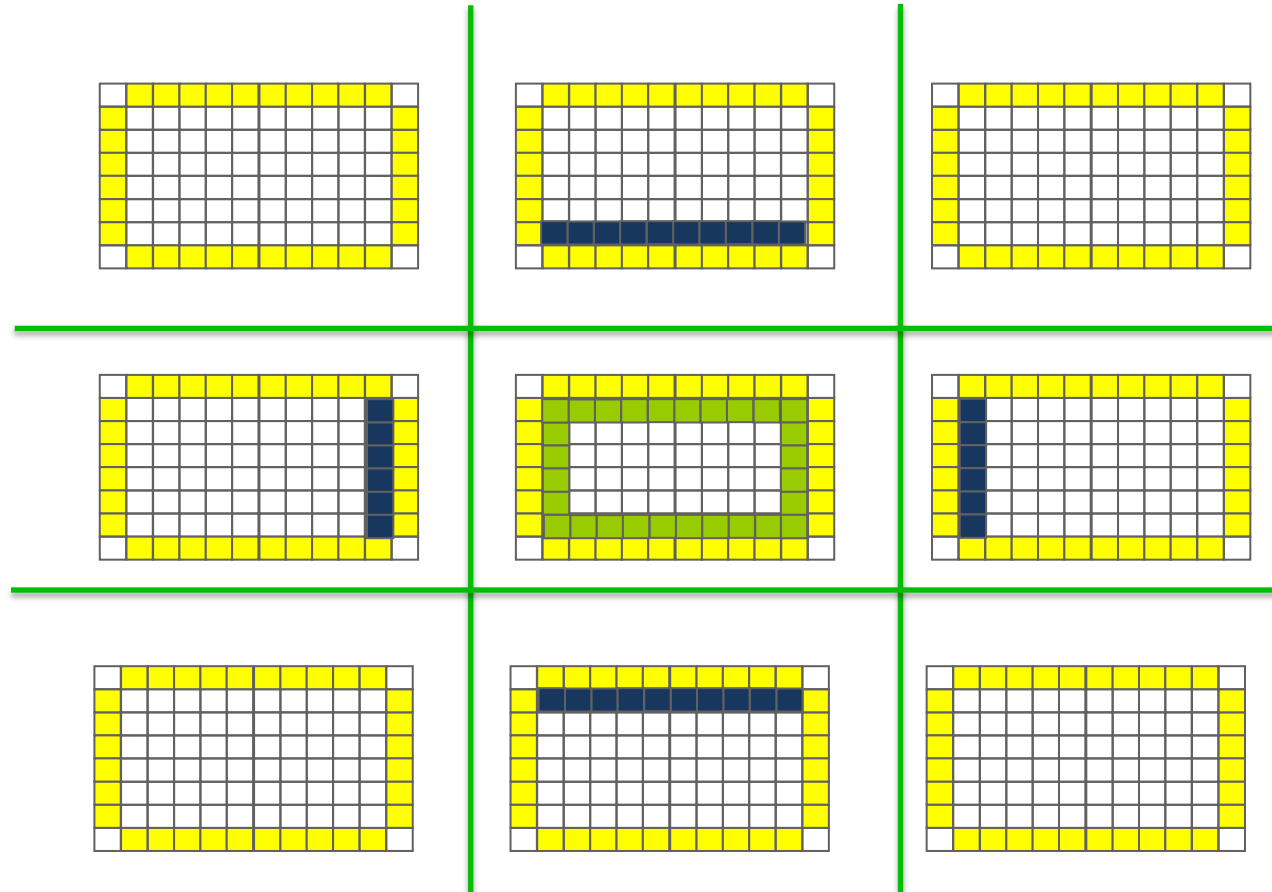
- Each process has its local “patch” of the global array
 - “bx” and “by” are the sizes of the local array
 - Always allocate a halo around the patch
 - Array allocated of size $(bx+2) \times (by+2)$



Check the **alloc_bufs** function to see how buffers are allocated

Necessary Data Transfers

- Provide access to remote data through a halo exchange (5 point stencil)



Check the **update_grid** function to see how it is done

MPI Examples on Cooley

- `/lus/theta-fs0/projects/ATPESC2020/mpi`
- Copy to your own project directory
- Submit job using “ATPESC2020” project, and “training” queue to use the reservation

Example: Stencil with Nonblocking Send/recv

- *nonblocking_p2p/stencil.c*
- Simple stencil code using nonblocking point-to-point operations

```
# make  
# qsub -A ATPESC2020 -q training -n <nproc> -t 10 ./job.sh
```

```
# mpirun -n <nproc> ./stencil <domain size> <heat source value> <iterations> <px> <py>
```

Derived Datatypes



U.S. DEPARTMENT OF
ENERGY

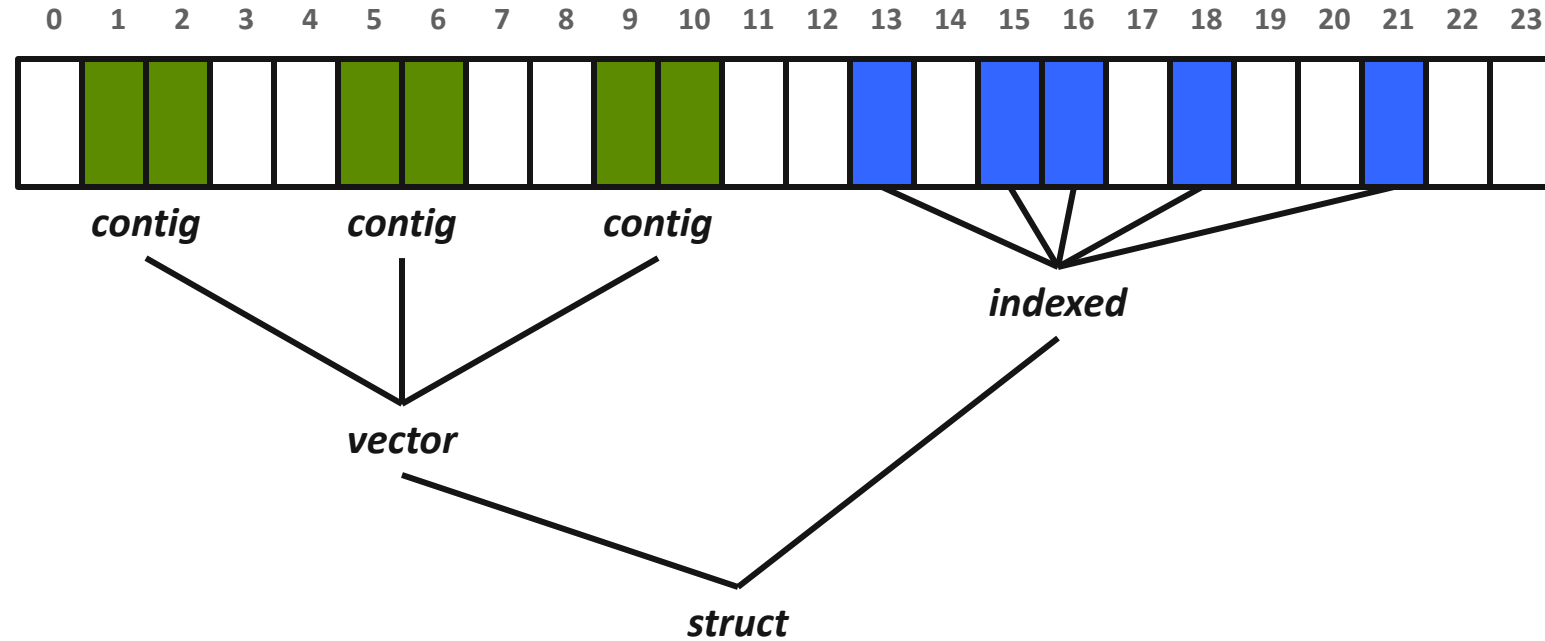
Introduction to Datatypes in MPI

- Datatypes allow to (de)serialize **arbitrary** data layouts into a message stream
 - Networks provide serial channels
 - Same for block devices and I/O
- Several constructors allow arbitrary layouts
 - Recursive specification possible
 - *Declarative* specification of data-layout
 - “what” and not “how”, leaves optimization to implementation (*many **unexplored** possibilities!*)
 - Choosing the right constructors is not always simple

Simple/Predefined Datatypes

- Equivalents exist for all C, C++ and Fortran native datatypes
 - C int → MPI_INT
 - C float → MPI_FLOAT
 - C double → MPI_DOUBLE
 - C uint32_t → MPI_UINT32_T
 - Fortran integer → MPI_INTEGER
- For more complex or user-created datatypes, MPI provides routines to represent them as well
 - Contiguous
 - Vector/Hvector
 - Indexed/Indexed_block/Hindexed/Hindexed_block
 - Struct
 - Some convenience types (e.g., subarray)

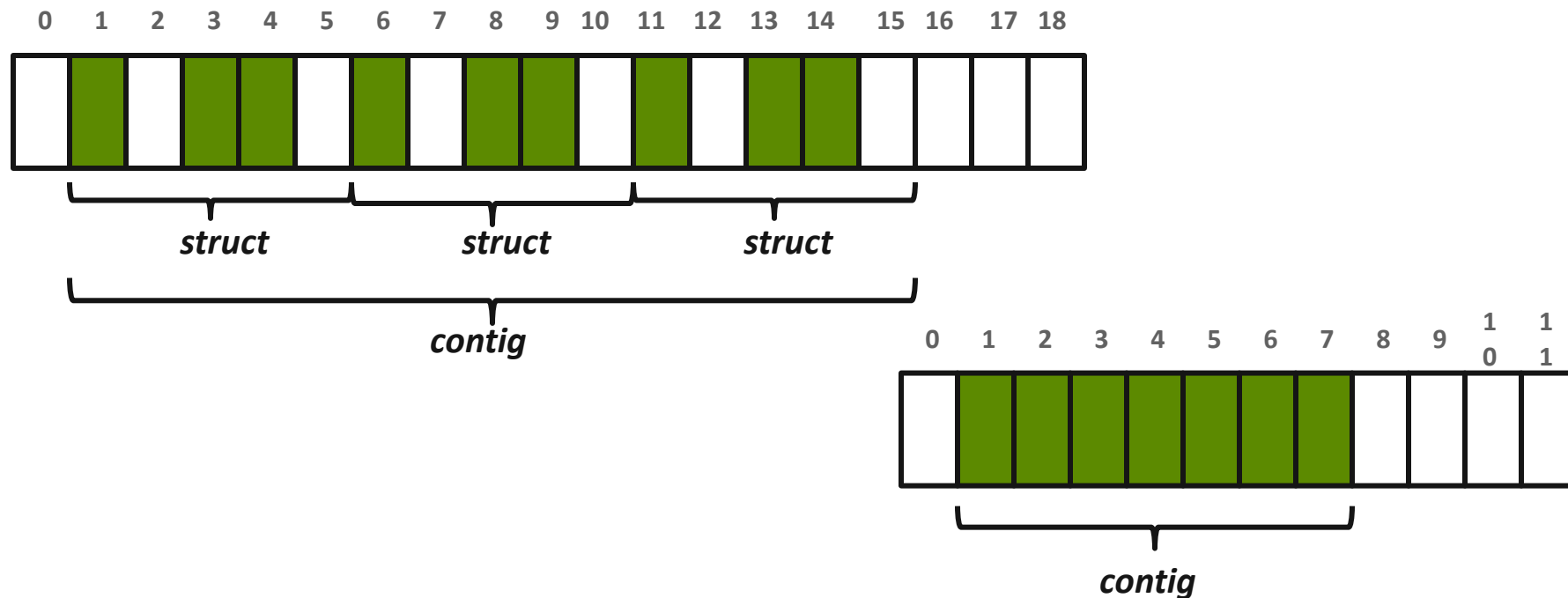
Derived Datatype Example



MPI_Type_contiguous

```
MPI_Type_contiguous(int count, MPI_Datatype oldtype, MPI_Datatype *newtype)
```

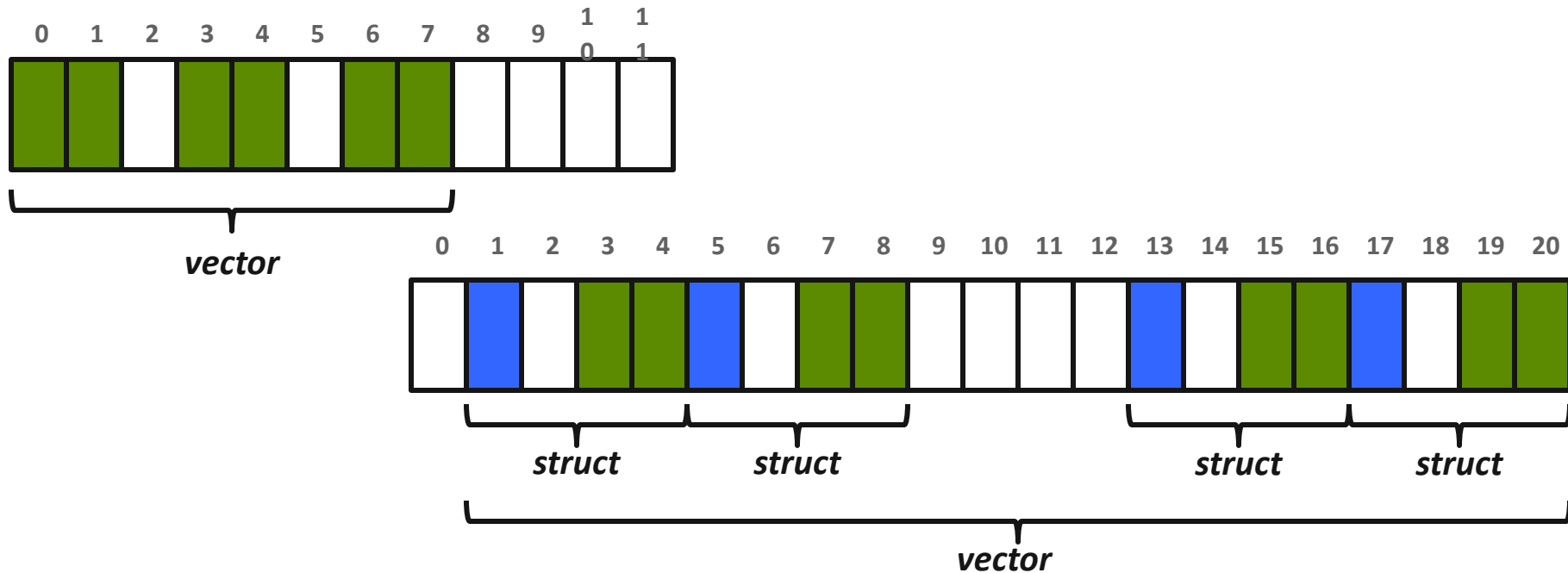
- Contiguous array of oldtype
- Should not be used as last type (can be replaced by count)



MPI_Type_vector

```
MPI_Type_vector(int count, int blocklen, int stride, MPI_Datatype oldtype,  
                MPI_Datatype *newtype)
```

- Specify strided blocks of data of oldtype
- Very useful for Cartesian arrays



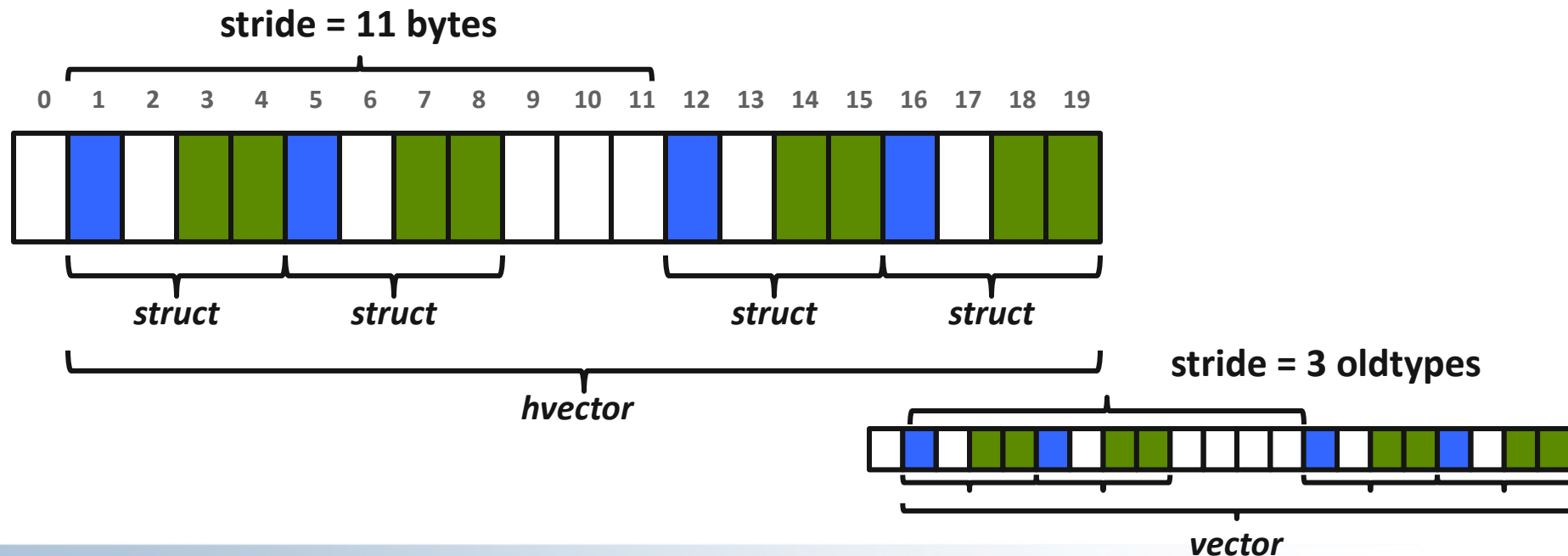
Commit, Free, and Dup

- Types must be committed before use
 - Only the ones that are used!
 - MPI_Type_commit may perform heavy optimizations (and will hopefully)
- MPI_Type_free
 - Free MPI resources of datatypes
 - Does not affect types built from it
- MPI_Type_dup
 - Duplicates a type
 - Library abstraction (composability)

MPI_Type_create_hvector

```
MPI_Type_create_hvector(int count, int blocklen, MPI_Aint stride, MPI_Datatype oldtype,  
                        MPI_Datatype *newtype)
```

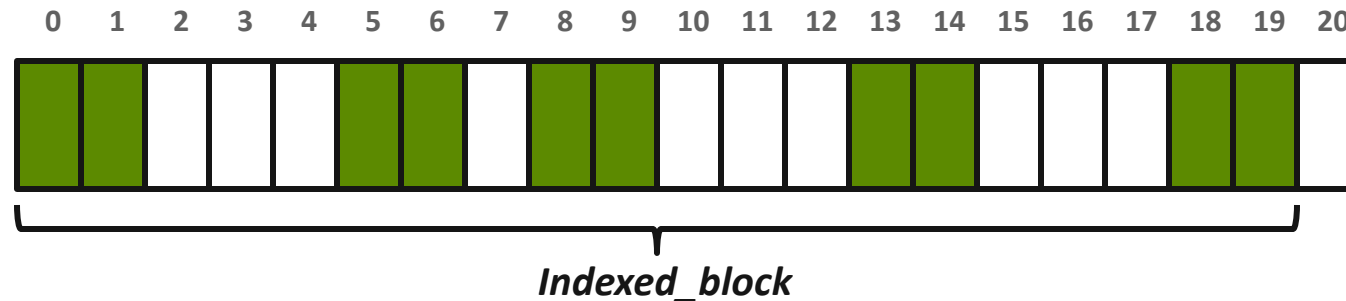
- Create byte strided vectors
- Useful for composition, e.g., vector of structs



MPI_Type_create_indexed_block

```
MPI_Type_create_indexed_block(int count, int blocklen, int *array_of_displacements,  
                             MPI_Datatype oldtype, MPI_Datatype *newtype)
```

- Pulling irregular subsets of data from a single array
 - dynamic codes with index lists, expensive though!
 - blen=2
 - displs={0,5,8,13,18}



MPI_Type_indexed

```
MPI_Type_indexed(int count, int* array_of_blocklens, int *array_of_displacements,  
MPI_Datatype oldtype, MPI_Datatype *newtype)
```

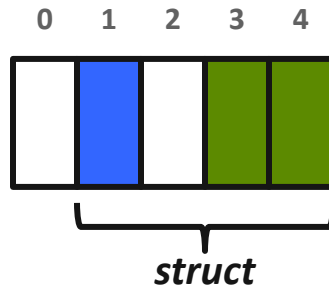
- Like indexed_block, but can have different block lengths
 - blen={1,1,2,1,2,1}
 - displs={0,3,5,9,13,17}



MPI_Type_create_struct

```
MPI_Type_create_struct(int count, int *array_of_blocklens, int *array_of_displacements,  
                      MPI_Datatype *array_of_types, MPI_Datatype *newtype)
```

- Most general constructor, allows different types and arbitrary arrays (also most costly)



MPI_Type_create_subarray

```
MPI_Type_create_subarray(int ndims, int* array_of_sizes, int *array_of_subsizes,  
                        int *array_of_starts, int order, MPI_Datatype oldtype,  
                        MPI_Datatype *newtype)
```

- Convenience function for creating datatypes for array segments
- Specify subarray of n-dimensional array (sizes) by start (starts) and size (subsize)

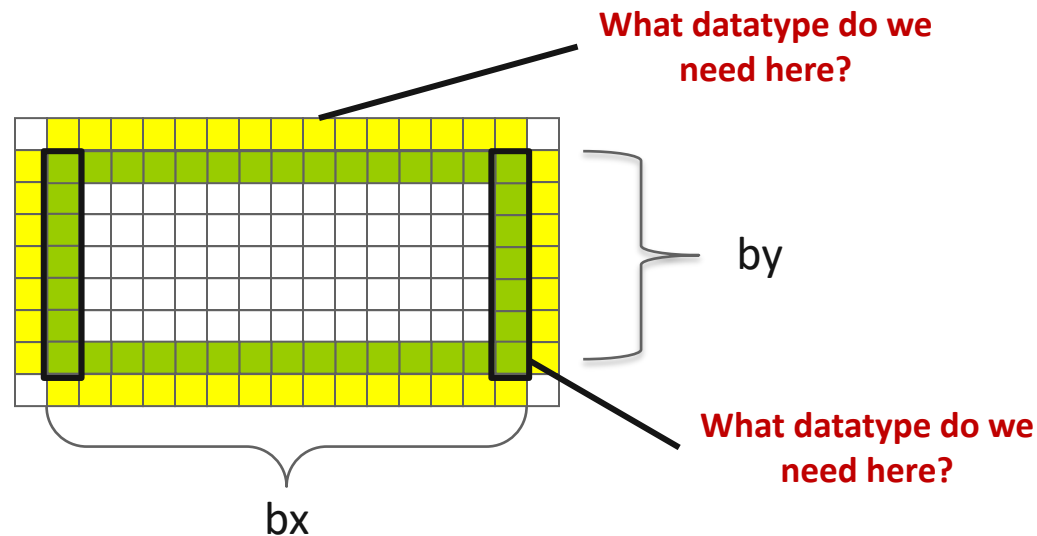
(0,0)	(0,1)	(0,2)	(0,3)
(1,0)	(1,1)	(1,2)	(1,3)
(2,0)	(2,1)	(2,2)	(2,3)
(3,0)	(3,1)	(3,2)	(3,3)

Section Summary

- Derived datatypes are a sophisticated mechanism to describe ANY layout in memory
 - Hierarchical construction of derived datatypes allows them to be just as complex as the data layout is
 - More complex layouts require more complex datatype constructions
- Current state of MPI implementations might be a bit lagging in performance, but it is improving
 - Increasing amount of hardware support to process derived datatypes on the network hardware
 - If the performance is lagging when you try it out, complain to the MPI implementer, don't just stop using it!

Exercise: Stencil with Derived Datatypes (1/2)

- In the basic version of the stencil code
 - Used nonblocking communication 👍
 - Used manual packing/unpacking of data 🙅
- Let's try to use derived datatypes
 - Specify the locations of the data instead of manually packing/unpacking



Exercise: Stencil with Derived Datatypes (2/2)

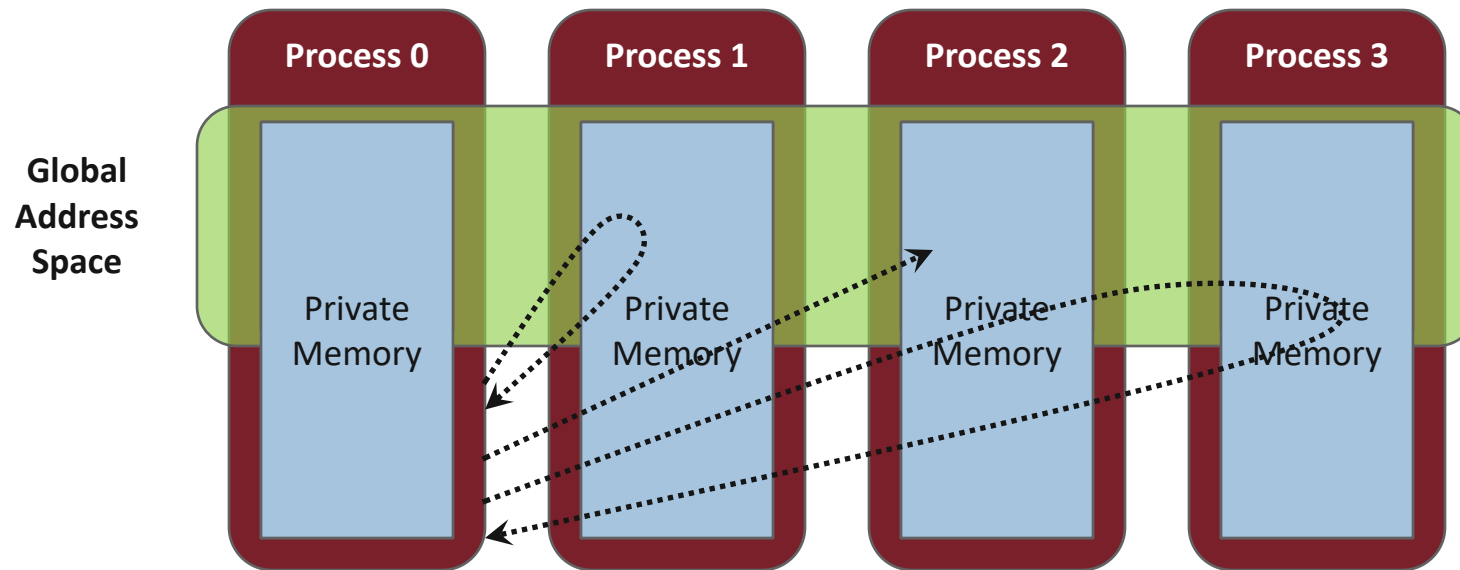
- Nonblocking sends and receives
- Data location specified by MPI datatypes
- Manual packing of data no longer required
- *Start from `nonblocking_p2p/stencil.c`*
- *Solution in `derived_datatype/stencil.c`*

MPI One-sided Communication

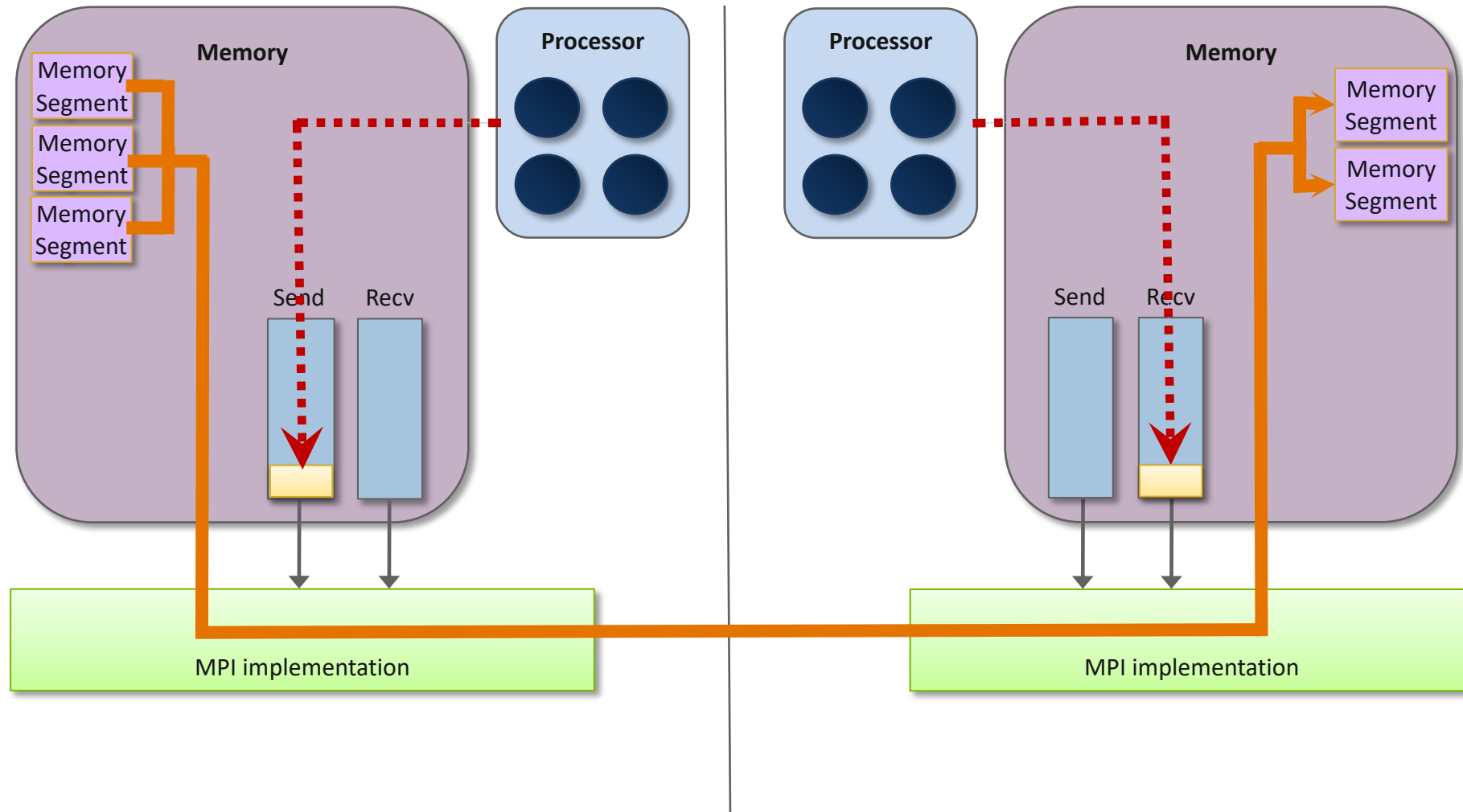


One-sided Communication

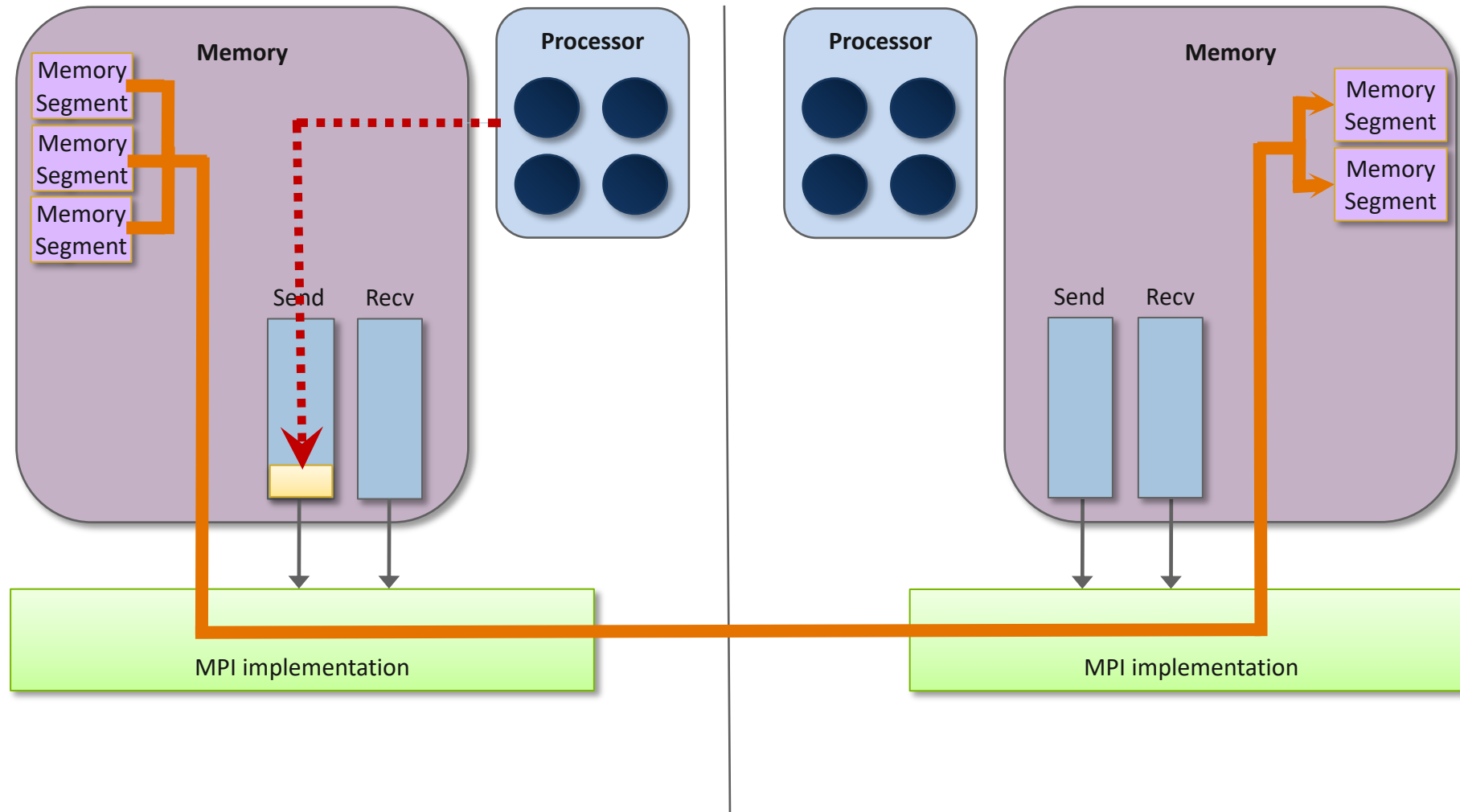
- The basic idea of one-sided communication models is to decouple data movement with process synchronization
 - Should be able to move data without requiring that the remote process synchronize
 - Each process exposes a part of its memory to other processes
 - Other processes can directly read from or write to this memory



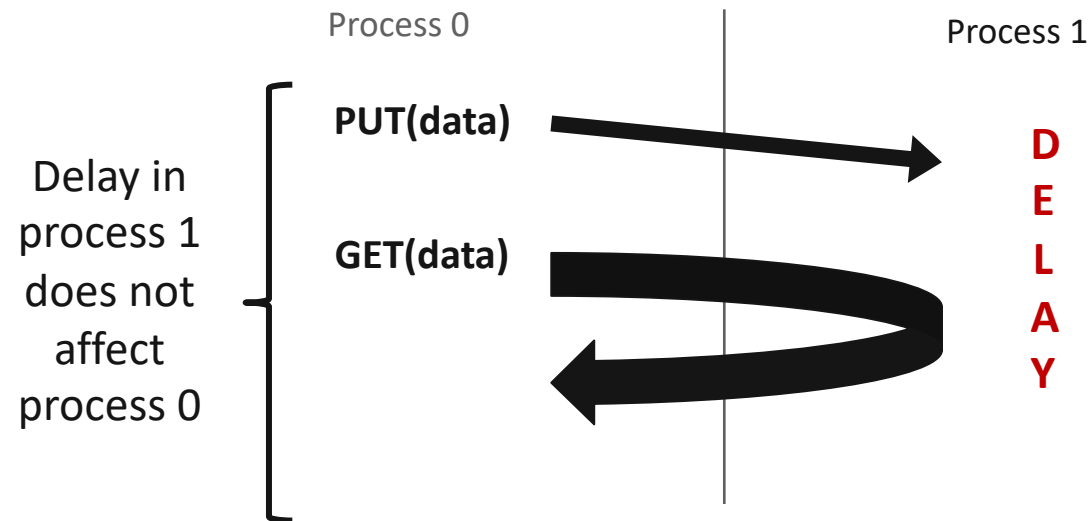
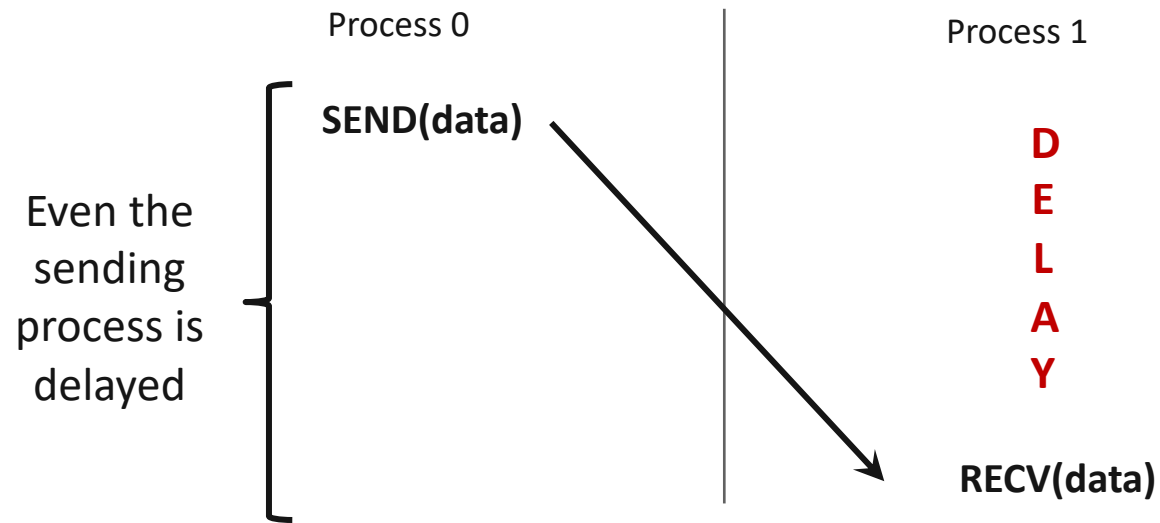
Two-sided Communication Example



One-sided Communication Example



Comparing One-sided and Two-sided Programming



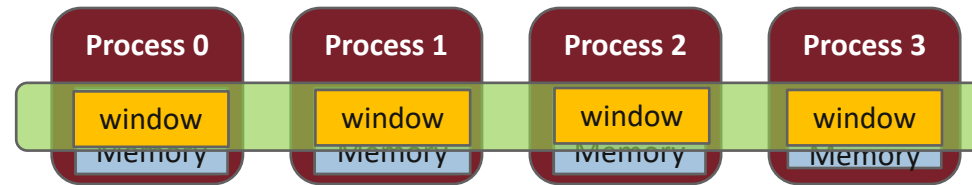
What we need to know in MPI RMA

- How to create remote accessible memory?
- Reading, Writing and Updating remote memory
- Data Synchronization
- Memory Model

Creating Public Memory

- Any memory used by a process is, by default, only locally accessible

- `X = malloc(100);`



- Once the memory is allocated, the user has to make an explicit MPI call to declare a memory region as remotely accessible
 - MPI terminology for remotely accessible memory is a “**window**”
 - A group of processes collectively create a “window”
- Once a memory region is declared as remotely accessible, all processes in the window can read/write data to this memory without explicitly synchronizing with the target process

Window creation models

- Four models exist
 - **MPI_WIN_ALLOCATE**
 - You want to create a buffer and directly make it remotely accessible
 - **MPI_WIN_CREATE**
 - You already have an allocated buffer that you would like to make remotely accessible
 - **MPI_WIN_CREATE_DYNAMIC**
 - You don't have a buffer yet, but will have one in the future
 - You may want to dynamically add/remove buffers to/from the window
 - **MPI_WIN_ALLOCATE_SHARED**
 - You want multiple processes on the same node share a buffer

MPI_WIN_ALLOCATE

```
MPI_Win_allocate(MPI_Aint size, int disp_unit, MPI_Info info, MPI_Comm comm,  
                 void *baseptr, MPI_Win *win)
```

- Create a remotely accessible memory region in an RMA window
 - Only data exposed in a window can be accessed with RMA ops.
- Arguments:
 - size - size of local data in bytes (nonnegative integer)
 - disp_unit - local unit size for displacements, in bytes (positive integer)
 - info - info argument (handle)
 - comm - communicator (handle)
 - baseptr - pointer to exposed local data
 - win - window (handle)

Example with MPI_WIN_ALLOCATE

```
int main(int argc, char ** argv)
{
    int *a;
    MPI_Win win;

    MPI_Init(&argc, &argv);

    /* collectively create remote accessible memory in a window */
    MPI_Win_allocate(1000*sizeof(int), sizeof(int), MPI_INFO_NULL, MPI_COMM_WORLD, &a, &win);

    /* Array 'a' is now accessible from all processes in
     * MPI_COMM_WORLD */

    MPI_Win_free(&win);

    MPI_Finalize(); return 0;
}
```

MPI_WIN_CREATE

```
MPI_Win_create(void *base, MPI_Aint size, int disp_unit, MPI_Info info, MPI_Comm comm,  
               MPI_Win *win)
```

- Expose a region of memory in an RMA window
 - Only data exposed in a window can be accessed with RMA ops.
- Arguments:
 - base - pointer to local data to expose
 - size - size of local data in bytes (nonnegative integer)
 - disp_unit - local unit size for displacements, in bytes (positive integer)
 - info - info argument (handle)
 - comm - communicator (handle)
 - win - window (handle)

Example with MPI_WIN_CREATE

```
int main(int argc, char ** argv)
{
    int *a;    MPI_Win win;

    MPI_Init(&argc, &argv);

    /* create private memory */
    a = (int *) malloc(1000*sizeof(int));
    /* use private memory like you normally would */
    for (int i = 0; i < 1000; i++) a[i] = i + 1;

    /* collectively declare memory as remotely accessible */
    MPI_Win_create(a, 1000*sizeof(int), sizeof(int), MPI_INFO_NULL, MPI_COMM_WORLD, &win);

    /* Array 'a' is now accessibly by all processes in
     * MPI_COMM_WORLD */

    MPI_Win_free(&win);
    free(a);
    MPI_Finalize(); return 0;
}
```

MPI_WIN_CREATE_DYNAMIC

```
MPI_Win_create_dynamic(MPI_Info info, MPI_Comm comm, MPI_Win *win)
```

- Create an RMA window, to which data can later be attached
 - Only data exposed in a window can be accessed with RMA ops
- Initially “**empty**”
 - Application can dynamically attach/detach memory to this window by calling **MPI_Win_attach/detach**
 - Application can access data on this window only after a memory region has been attached
- Window origin is **MPI_BOTTOM**
 - Displacements are segment addresses relative to **MPI_BOTTOM**
 - Must tell others the displacement after calling attach

Example with MPI_WIN_CREATE_DYNAMIC

```
int main(int argc, char ** argv)
{
    int *a;    MPI_Win win;

    MPI_Init(&argc, &argv);
    MPI_Win_create_dynamic(MPI_INFO_NULL, MPI_COMM_WORLD, &win);

    /* create private memory */
    a = (int *) malloc(1000 * sizeof(int));
    /* use private memory like you normally would */
    for (int i = 0; i < 1000; i++) a[i] = i + 1;

    /* locally declare memory as remotely accessible */
    MPI_Win_attach(win, a, 1000*sizeof(int));

    /* Array 'a' is now accessible from all processes */

    /* undeclare remotely accessible memory */
    MPI_Win_detach(win, a); free(a);
    MPI_Win_free(&win);

    MPI_Finalize(); return 0;
}
```

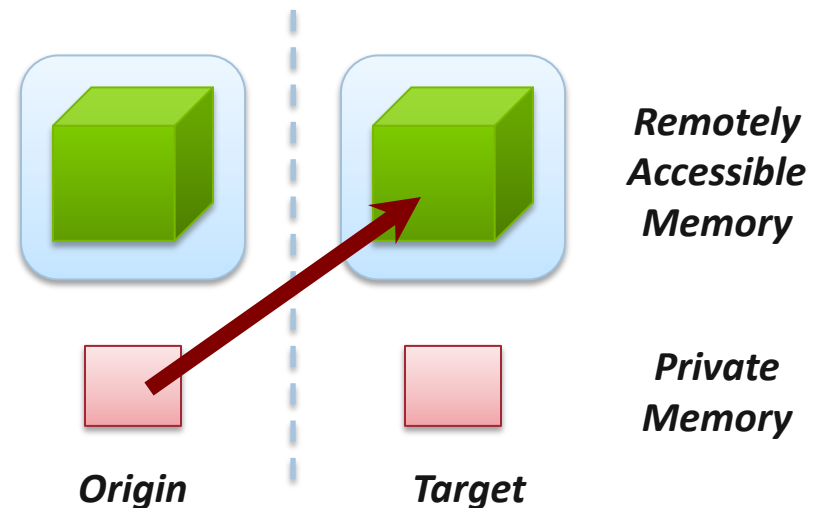

Data movement

- MPI provides ability to read, write and atomically modify data in remotely accessible memory regions
 - `MPI_PUT`
 - `MPI_GET`
 - `MPI_ACCUMULATE` `(atomic)`
 - `MPI_GET_ACCUMULATE` `(atomic)`
 - `MPI_COMPARE_AND_SWAP` `(atomic)`
 - `MPI_FETCH_AND_OP` `(atomic)`

Data movement: *Put*

```
MPI_Put(const void *origin_addr, int origin_count, MPI_Datatype origin_dtype,  
        int target_rank, MPI_Aint target_disp, int target_count,  
        MPI_Datatype target_dtype, MPI_Win win)
```

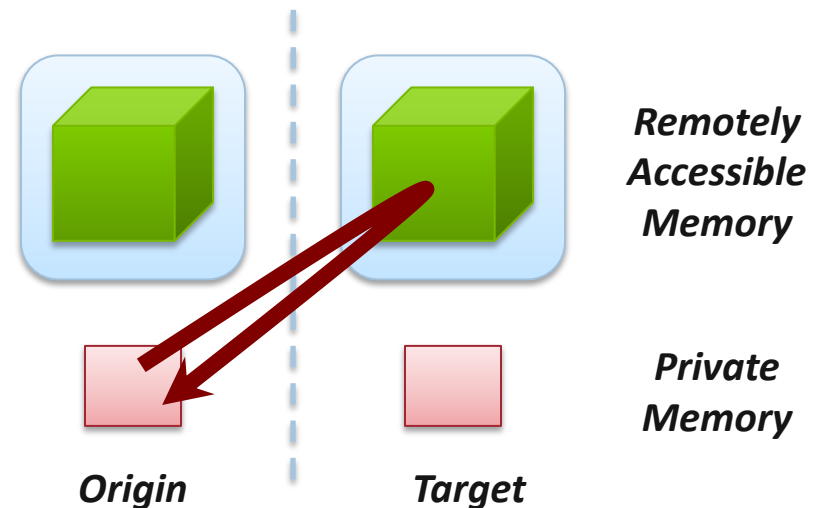
- Move data from origin, to target
- Separate data description triples for **origin** and **target**



Data movement: *Get*

```
MPI_Get(void *origin_addr, int origin_count, MPI_Datatype origin_dtype,  
        int target_rank, MPI_Aint target_disp, int target_count,  
        MPI_Datatype target_dtype, MPI_Win win)
```

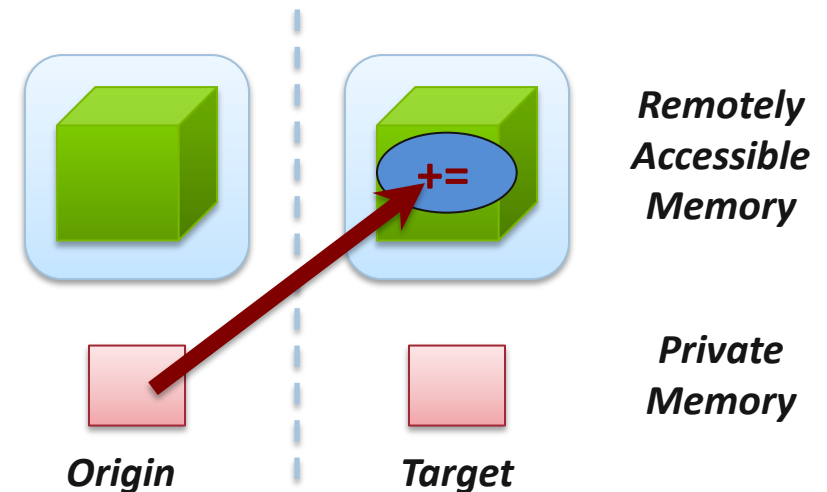
- Move data to origin, from target
- Separate data description triples for **origin** and **target**



Atomic Data Aggregation: *Accumulate*

```
MPI_Accumulate(const void *origin_addr, int origin_count, MPI_Datatype origin_dtype,  
              int target_rank, MPI_Aint target_disp, int target_count,  
              MPI_Datatype target_dtype, MPI_Op op, MPI_Win win)
```

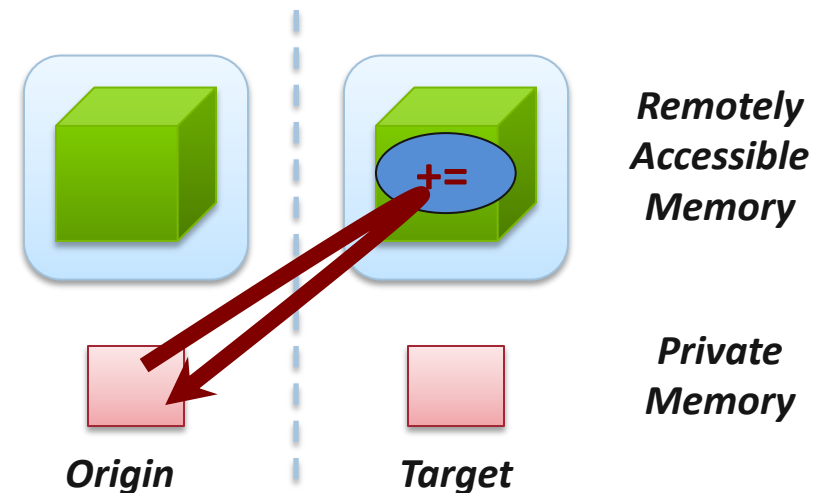
- Atomic update operation, similar to a put
 - Reduces origin and target data into target buffer using op argument as combiner
 - Op = **MPI_SUM**, **MPI_PROD**, **MPI_OR**, **MPI_REPLACE**, **MPI_NO_OP**, ...
 - Predefined ops only, no user-defined operations
- Different data layouts between target/origin OK
 - Basic type elements must match
- Op = **MPI_REPLACE**
 - Implements $f(a,b)=b$
 - Atomic PUT



Atomic Data Aggregation: *Get Accumulate*

```
MPI_Get_accumulate(const void *origin_addr, int origin_count, MPI_Datatype origin_dtype,  
                  void *result_addr, int result_count, MPI_Datatype result_dtype,  
                  int target_rank, MPI_Aint target_disp, int target_count,  
                  MPI_Datatype target_dtype, MPI_Op op, MPI_Win win)
```

- Atomic read-modify-write
 - Op = **MPI_SUM**, **MPI_PROD**, **MPI_OR**, **MPI_REPLACE**, **MPI_NO_OP**, ...
 - Predefined ops only
- Result stored in target buffer
- Original data stored in result buf
- Different data layouts between target/origin OK
 - Basic type elements must match
- Atomic get with **MPI_NO_OP**
- Atomic swap with **MPI_REPLACE**



Atomic Data Aggregation: *FOP* and *CAS*

```
MPI_Fetch_and_op(const void *origin_addr, void *result_addr, MPI_Datatype dtype,  
                int target_rank, MPI_Aint target_disp, MPI_Op op, MPI_Win win)
```

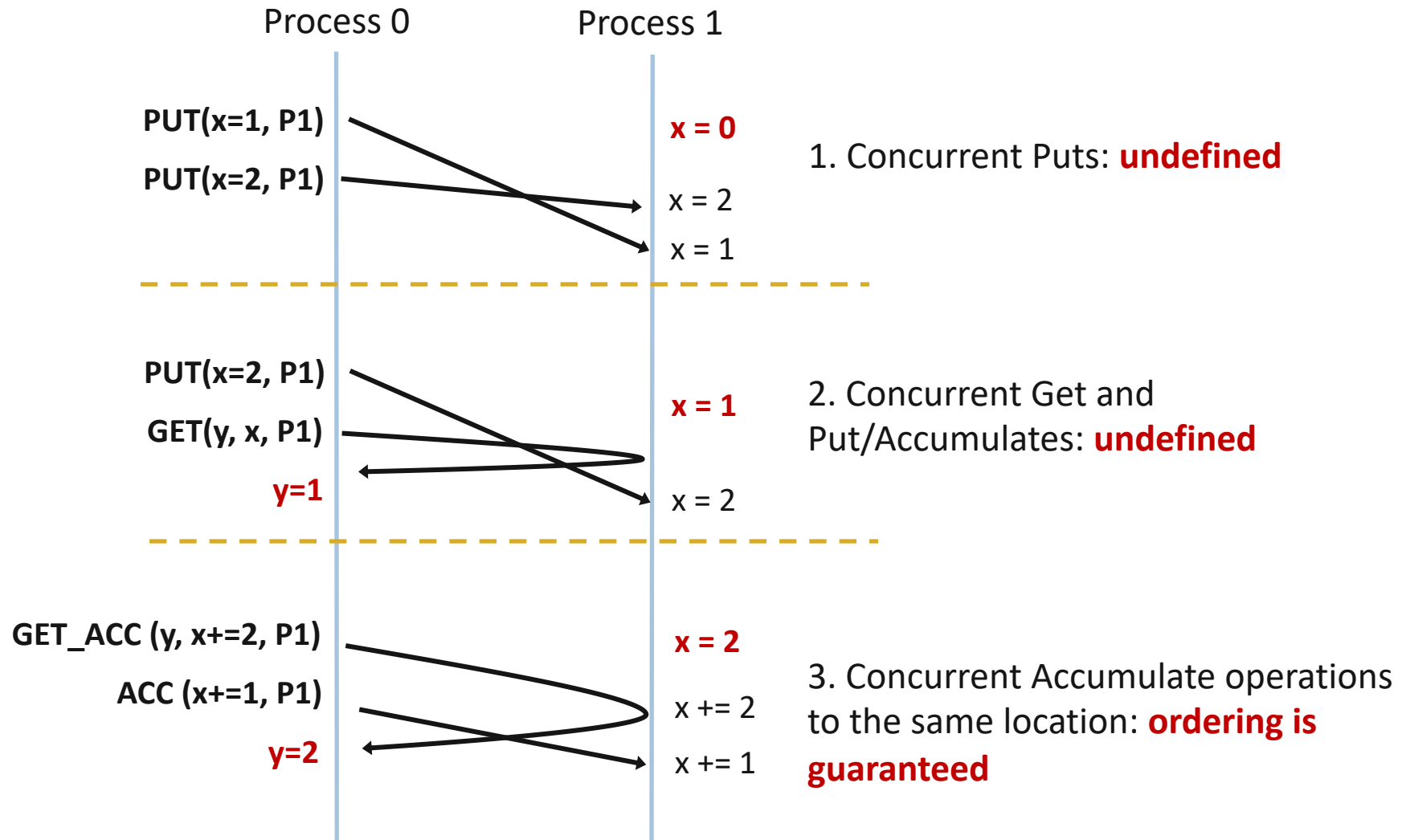
```
MPI_Compare_and_swap(const void *origin_addr, const void *compare_addr,  
                    void *result_addr, MPI_Datatype dtype, int target_rank,  
                    MPI_Aint target_disp, MPI_Win win)
```

- FOP: Simpler version of MPI_Get_accumulate
 - All buffers share a single predefined datatype
 - No count argument (it's always 1)
 - Simpler interface allows hardware optimization
- CAS: Atomic swap if target value is equal to compare value

Ordering of Operations in MPI RMA

- No guaranteed ordering for Put/Get operations
- Result of concurrent Puts to the same location undefined
- Result of Get concurrent Put/Accumulate undefined
 - Can be garbage in both cases
- Result of concurrent accumulate operations to the same location are defined according to the order in which they occurred
 - Atomic put: Accumulate with op = MPI_REPLACE
 - Atomic get: Get_accumulate with op = MPI_NO_OP
- Accumulate operations from a given process are ordered by default
 - User can tell the MPI implementation that (s)he does not require ordering as optimization hint
 - You can ask for only the needed orderings: RAW (read-after-write), WAR, RAR, or WAW

Examples with operation ordering



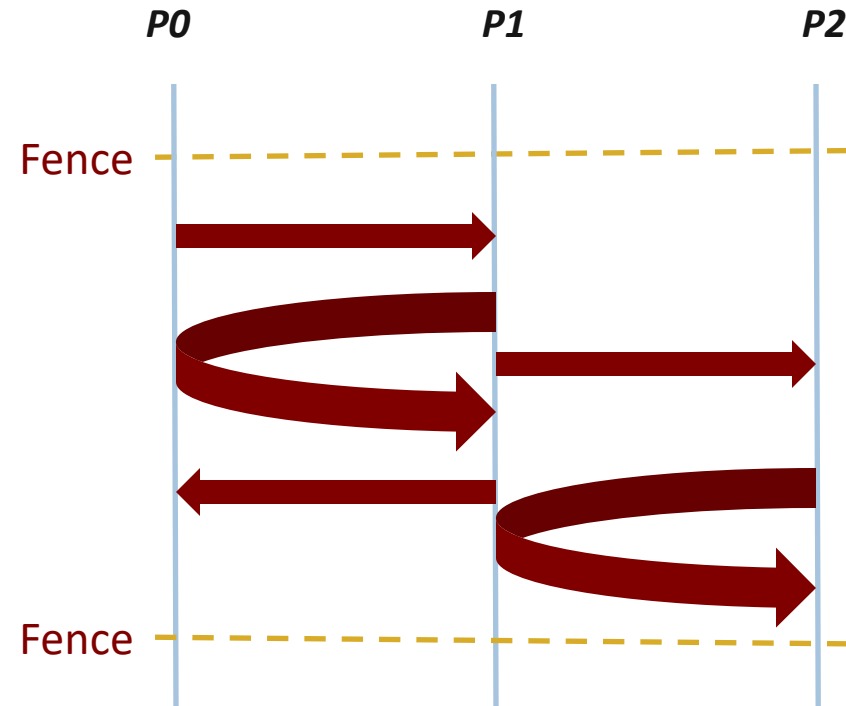
RMA Synchronization Models

- RMA data access model
 - When is a process allowed to read/write remotely accessible memory?
 - When is data written by process X is available for process Y to read?
 - RMA synchronization models define these semantics
- Three synchronization models provided by MPI:
 - Fence (active target)
 - Post-start-complete-wait (generalized active target)
 - Lock/Unlock (passive target)
- Data accesses occur within “epochs”
 - *Access epochs*: contain a set of operations issued by an origin process
 - *Exposure epochs*: enable remote processes to update a target’s window
 - Epochs define ordering and completion semantics
 - Synchronization models provide mechanisms for establishing epochs
 - E.g., starting, ending, and synchronizing epochs

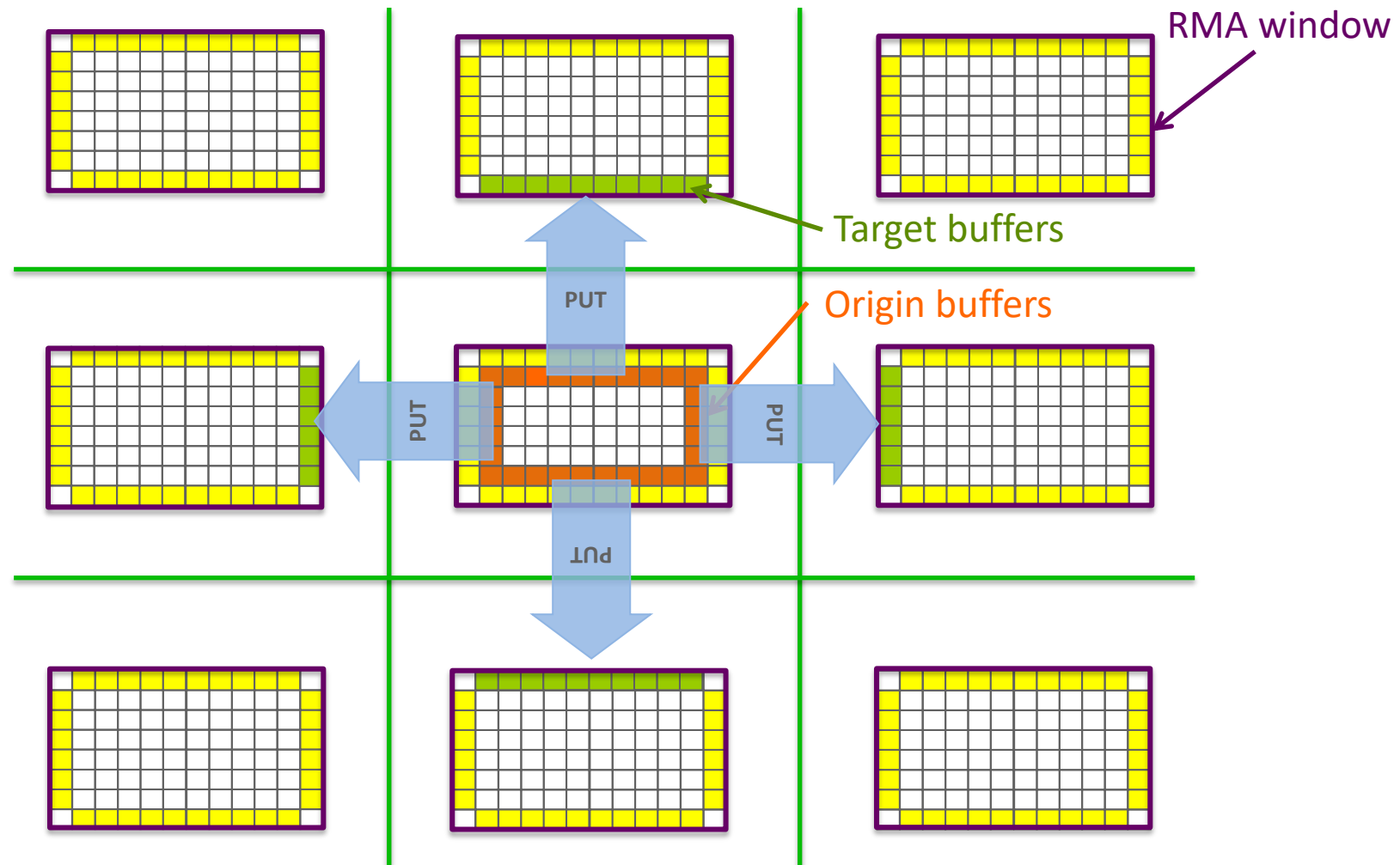
Fence: Active Target Synchronization

```
MPI_Win_fence(int assert, MPI_Win win)
```

- Collective synchronization model
- Starts *and* ends access and exposure epochs on all processes in the window
- All processes in group of “win” do an **MPI_WIN_FENCE** to open an epoch
- Everyone can issue **PUT/GET** operations to read/write data
- Everyone does an **MPI_WIN_FENCE** to close the epoch
- All operations complete at the second fence synchronization



Exercise 1: Stencil with RMA Fence (1/2)



Exercise 1: Stencil with RMA Fence (2/2)

- In the derived datatype version of the stencil code
 - Used nonblocking communication
 - Used derived datatypes
- Let's try to use RMA fence
 - Move data with PUT instead of send/recv
- *Start from `derived_datatype/stencil.c`*
- *Solution available in `rma/stencil_fence_put.c`*

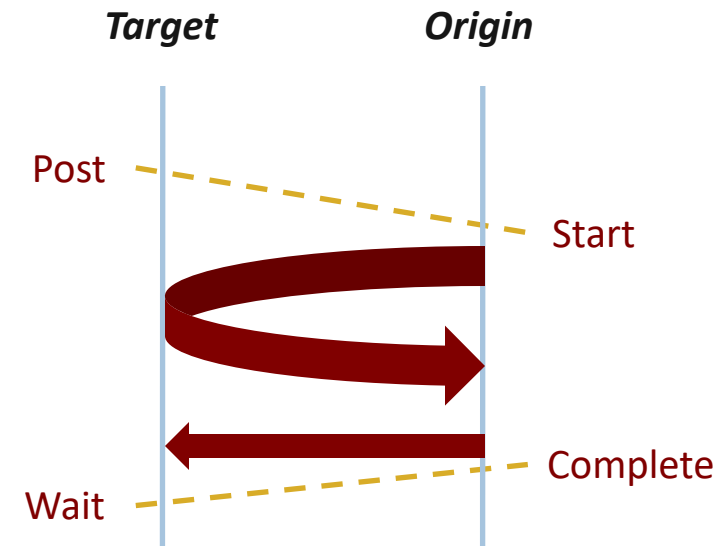
Exercise 2: Stencil with RMA Fence (GET model)

- In the derived datatype version of the stencil code
 - Used nonblocking communication
 - Used derived datatypes
- Let's try to use RMA fence
 - Move data with GET instead of send/recv
- *Start from `rma/stencil_fence_put.c`*
- *Solution available in `rma/stencil_fence_get.c`*

PSCW: Generalized Active Target Synchronization

```
MPI_Win_post/start(MPI_Group grp, int assert, MPI_Win win)
MPI_Win_complete/wait(MPI_Win win)
```

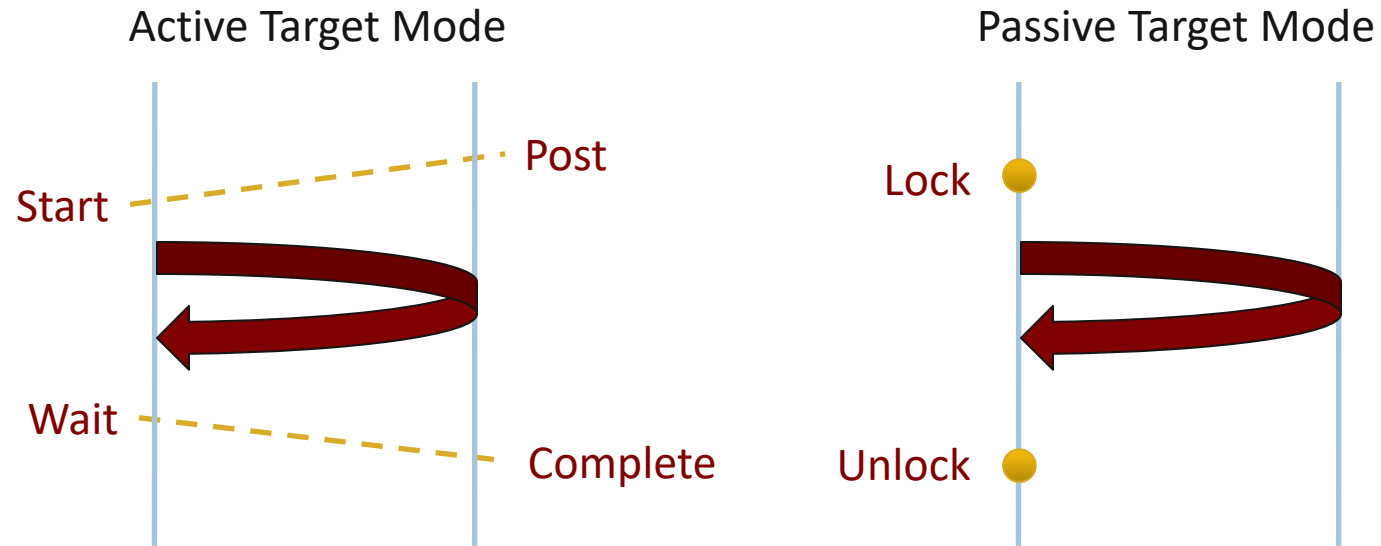
- Like FENCE, but origin and target specify who they communicate with
- Target: Exposure epoch
 - Opened with `MPI_Win_post`
 - Closed by `MPI_Win_wait`
- Origin: Access epoch
 - Opened by `MPI_Win_start`
 - Closed by `MPI_Win_complete`
- All synchronization operations may block, to enforce P-S/C-W ordering
 - Processes can be both origins and targets



Exercise 3: Stencil with RMA PSCW (PUT model)

- In the fence version of the stencil code
 - Unnecessary synchronization between all processes
- Let's try to use RMA PSCW
 - Synchronize with PSCW instead of Fence
- *Start from `rma/stencil_fence_put.c`*
- *Solution available in `rma/stencil_pscw_put.c`*

Lock/Unlock: Passive Target Synchronization



- Passive mode: One-sided, *asynchronous* communication
 - Target does **not** participate in communication operation
- Shared memory-like model

Passive Target Synchronization

```
MPI_Win_lock(int locktype, int rank, int assert, MPI_Win win)
```

```
MPI_Win_unlock(int rank, MPI_Win win)
```

```
MPI_Win_flush/flush_local(int rank, MPI_Win win)
```

- Lock/Unlock: Begin/end passive mode epoch
 - Target process does not make a corresponding MPI call
 - Can initiate multiple passive target epochs to different processes
 - Concurrent epochs to same process not allowed (affects threads)
- Lock type
 - SHARED: Other processes using shared can access concurrently
 - EXCLUSIVE: No other processes can access concurrently
- Flush: Remotely complete RMA operations to the target process
 - After completion, data can be read by target process or a different process
- Flush_local: Locally complete RMA operations to the target process

Advanced Passive Target Synchronization

```
MPI_Win_lock_all(int assert, MPI_Win win)
```

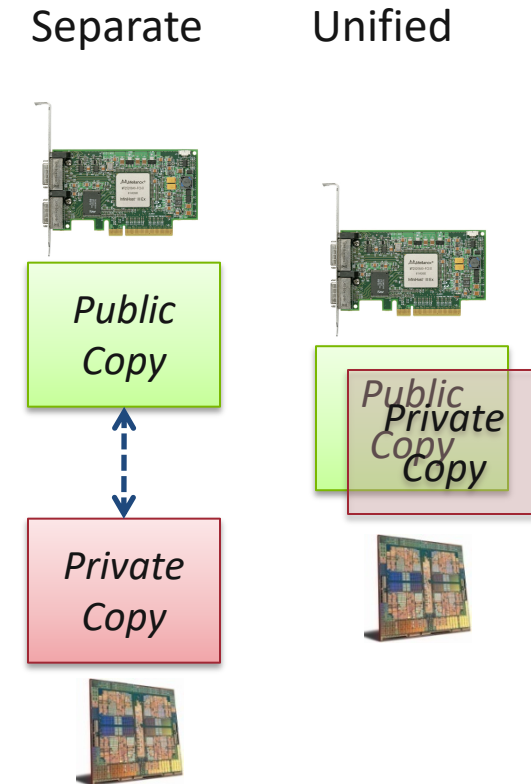
```
MPI_Win_unlock_all(MPI_Win win)
```

```
MPI_Win_flush_all/flush_local_all(MPI_Win win)
```

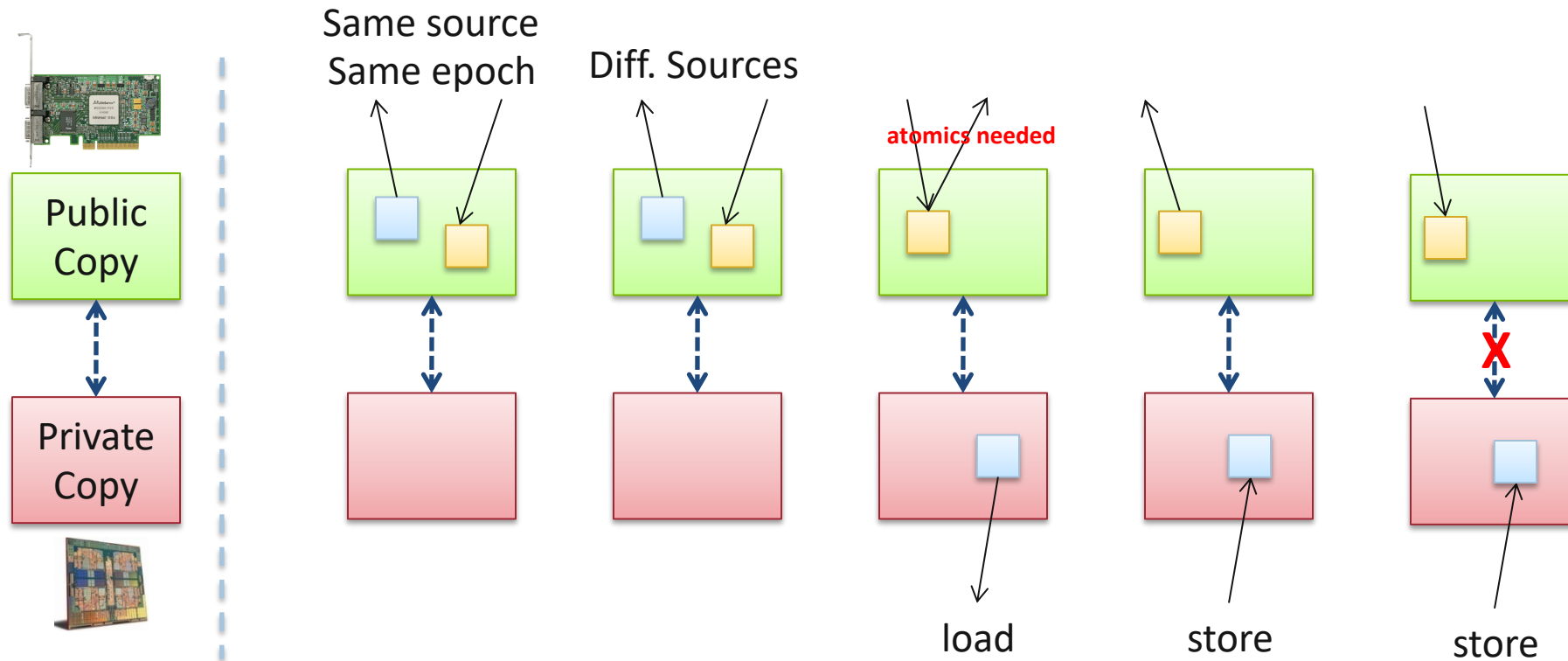
- **Lock_all**: Shared lock, passive target epoch to all other processes
 - Expected usage is long-lived: **lock_all**, **put/get**, **flush**, ..., **unlock_all**
- **Flush_all** – remotely complete RMA operations to all processes
- **Flush_local_all** – locally complete RMA operations to all processes

MPI RMA Memory Model

- MPI-3 provides two memory models: separate and unified
- Separate Model
 - Logical public and private copies
 - MPI provides software coherence between window copies
 - Extremely portable, to systems that don't provide hardware coherence
- New Unified Model
 - Single copy of the window
 - System must provide coherence
 - Superset of separate semantics
 - E.g. allows concurrent local/remote access
 - Provides access to full performance potential of hardware

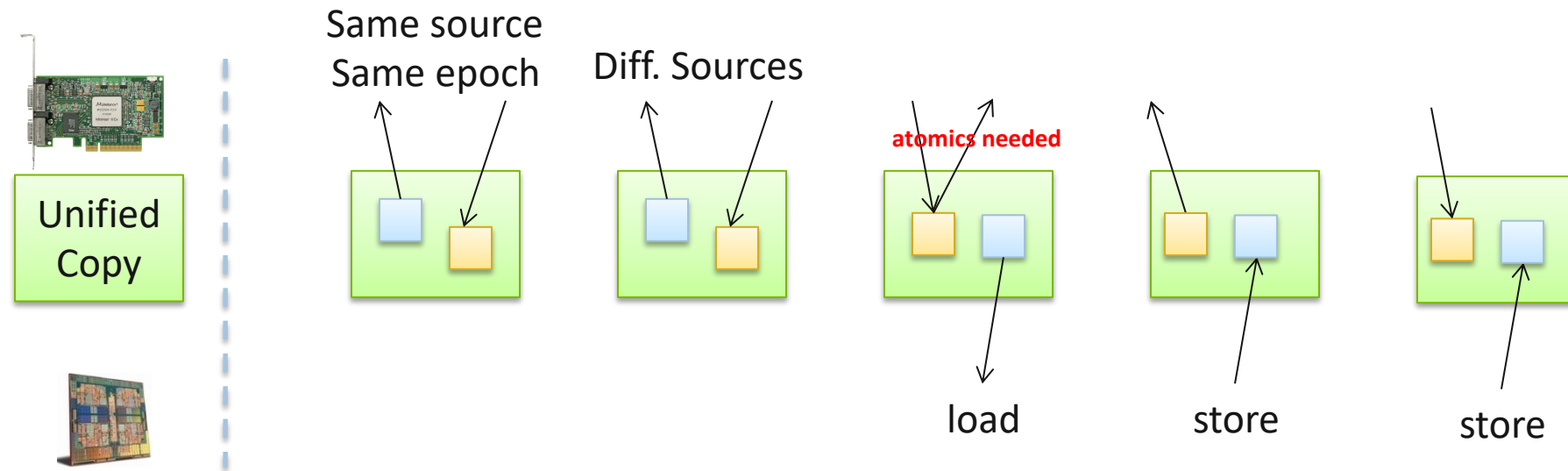


MPI RMA Memory Model (separate windows)



- Very portable, compatible with non-coherent memory systems
- Limits concurrent accesses to enable software coherence

MPI RMA Memory Model (unified windows)



- Allows concurrent local/remote accesses
- Concurrent, conflicting operations are allowed (not invalid)
 - Outcome is not defined by MPI (defined by the hardware)
- Can enable better performance by reducing synchronization

MPI RMA Operation Compatibility (Separate)

	Load	Store	Get	Put	Acc
Load	OVL+NOVL	OVL+NOVL	OVL+NOVL	NOVL	NOVL
Store	OVL+NOVL	OVL+NOVL	NOVL	X	X
Get	OVL+NOVL	NOVL	OVL+NOVL	NOVL	NOVL
Put	NOVL	X	NOVL	NOVL	NOVL
Acc	NOVL	X	NOVL	NOVL	OVL+NOVL

This matrix shows the compatibility of MPI-RMA operations when two or more processes access a window at the same target concurrently.

OVL – Overlapping operations permitted

NOVL – Nonoverlapping operations permitted

X – Combining these operations is OK, but data might be garbage

MPI RMA Operation Compatibility (Unified)

	Load	Store	Get	Put	Acc
Load	OVL+NOVL	OVL+NOVL	OVL+NOVL	NOVL	NOVL
Store	OVL+NOVL	OVL+NOVL	NOVL	NOVL	NOVL
Get	OVL+NOVL	NOVL	OVL+NOVL	NOVL	NOVL
Put	NOVL	NOVL	NOVL	NOVL	NOVL
Acc	NOVL	NOVL	NOVL	NOVL	OVL+NOVL

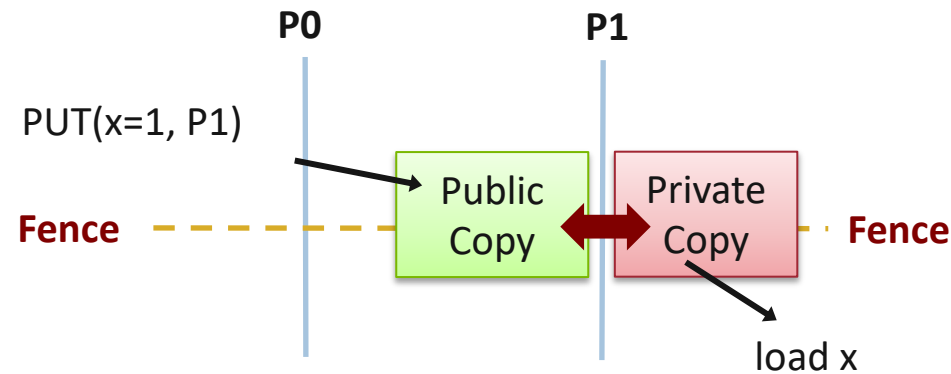
This matrix shows the compatibility of MPI-RMA operations when two or more processes access a window at the same target concurrently.

OVL – Overlapping operations permitted

NOVL – Nonoverlapping operations permitted

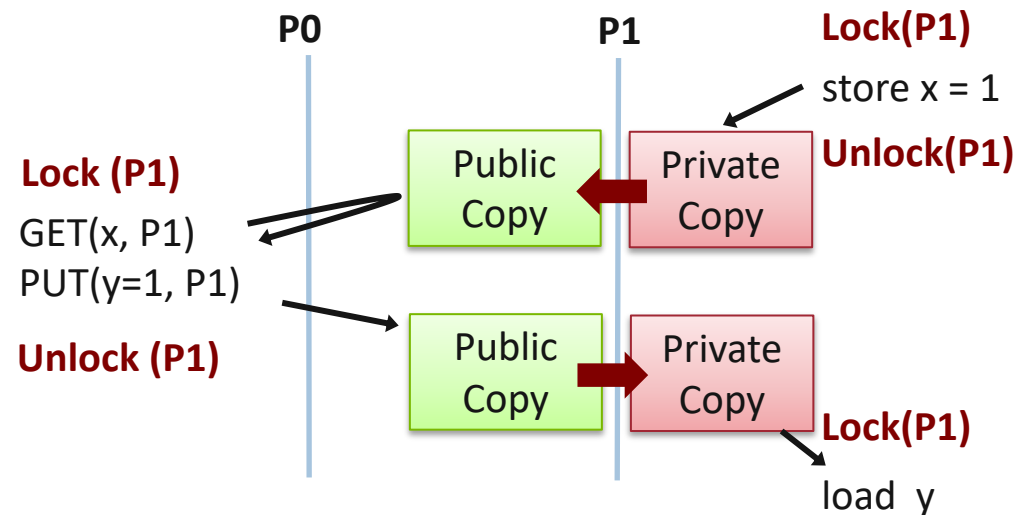
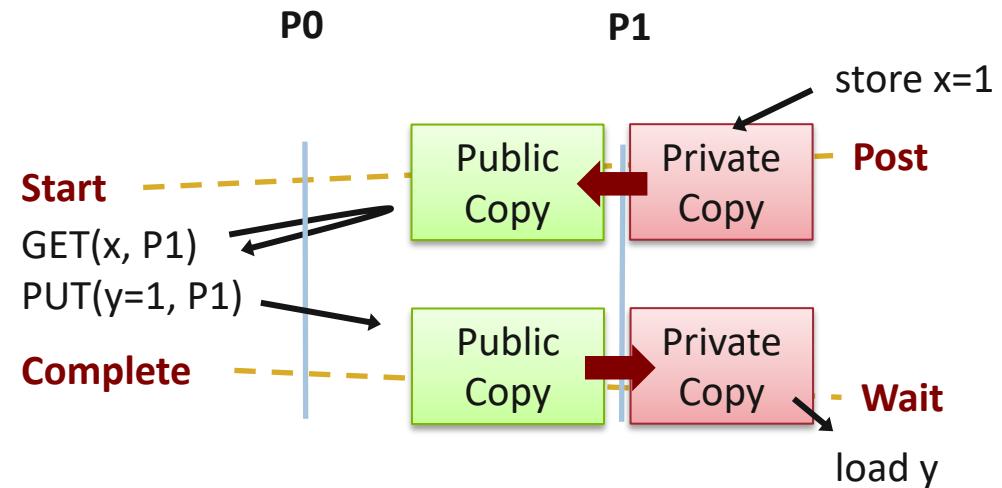
Synchronizing Local and RMA Access (1/2)

- RMA operations access the public copy of window
- Local load/store update the private copy
 - Including using as MPI send/receive buffers
- Ensure memory synchronization for portability
- Implicit memory synchronization (i.e., memory barrier) in RMA synchronization calls
 - **Fence:** Synchronize private and public copies of local window



Synchronizing Local and RMA Access (2/2)

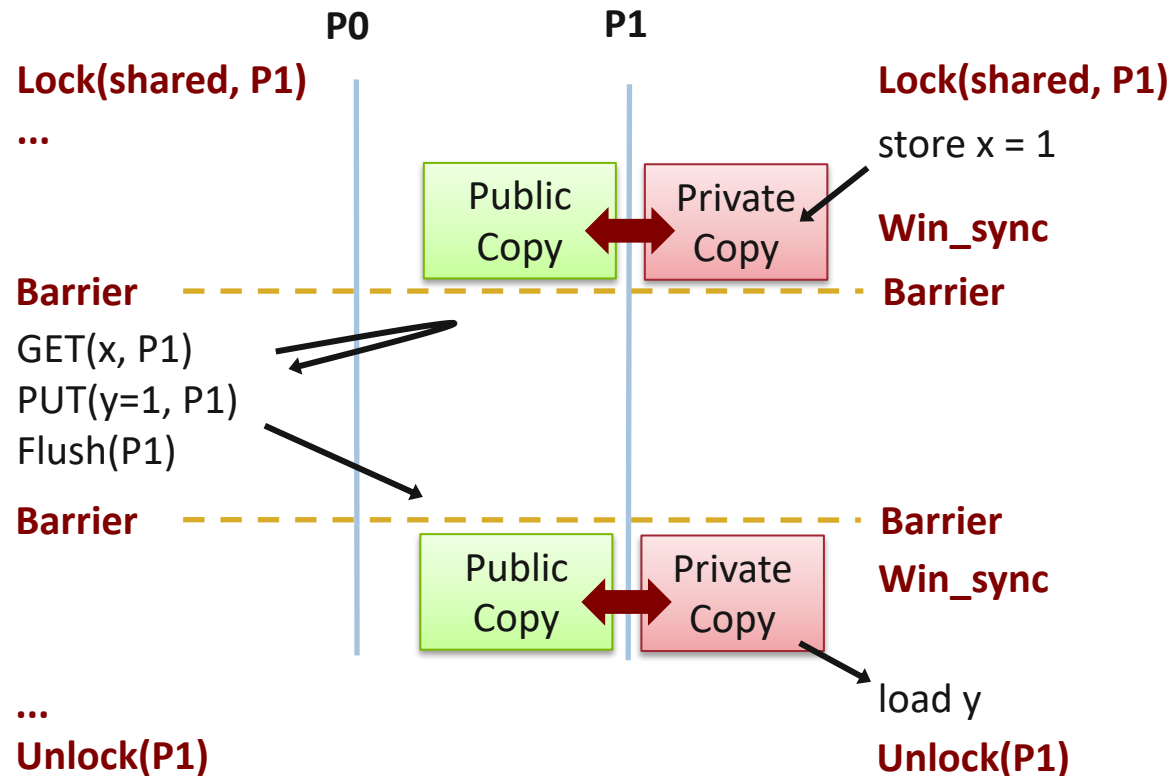
- PSCW active target epoch
 - **Post:** Updates in private copy becomes visible in public copy
 - **Wait:** Updates in public copy becomes visible in private copy
- Passive target epoch
 - **Lock/Lock_all:** Updates in public copy becomes visible in private
 - **Unlock/Unlock_all:** Updates in private copy becomes visible in public



Window synchronization: MPI_WIN_SYNC

```
MPI_Win_sync(MPI_Win win)
```

- Synchronizes the public and private copies of local window in passive target epoch



Exercise 4: Stencil with RMA Lock_all/Unlock_all (PUT model)

- In the fence and PSCW versions of the stencil code, RMA synchronization involves the target processes
- Let's try to use RMA Lock_all/Flush_all/Unlock_all
 - Only the origin processes call RMA synchronization
 - Still need **Barrier** for process synchronization (e.g., ensure neighbors have completed data update to my local window)
 - Need **Win_sync** for memory synchronization
- *Start from rma/stencil_fence_put.c*
- *Solution available in rma/stencil_lock_put.c*

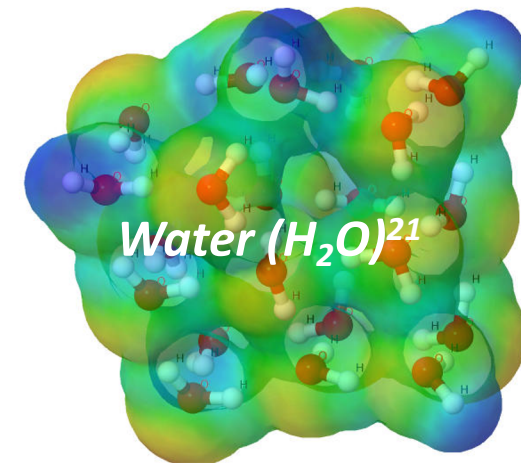
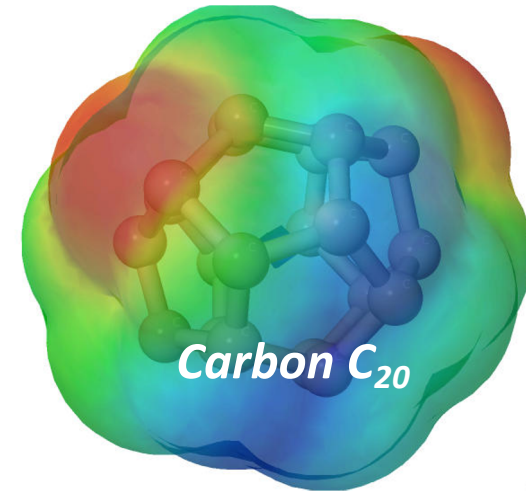
Example: Block Sparse Matrix Multiplication



U.S. DEPARTMENT OF
ENERGY

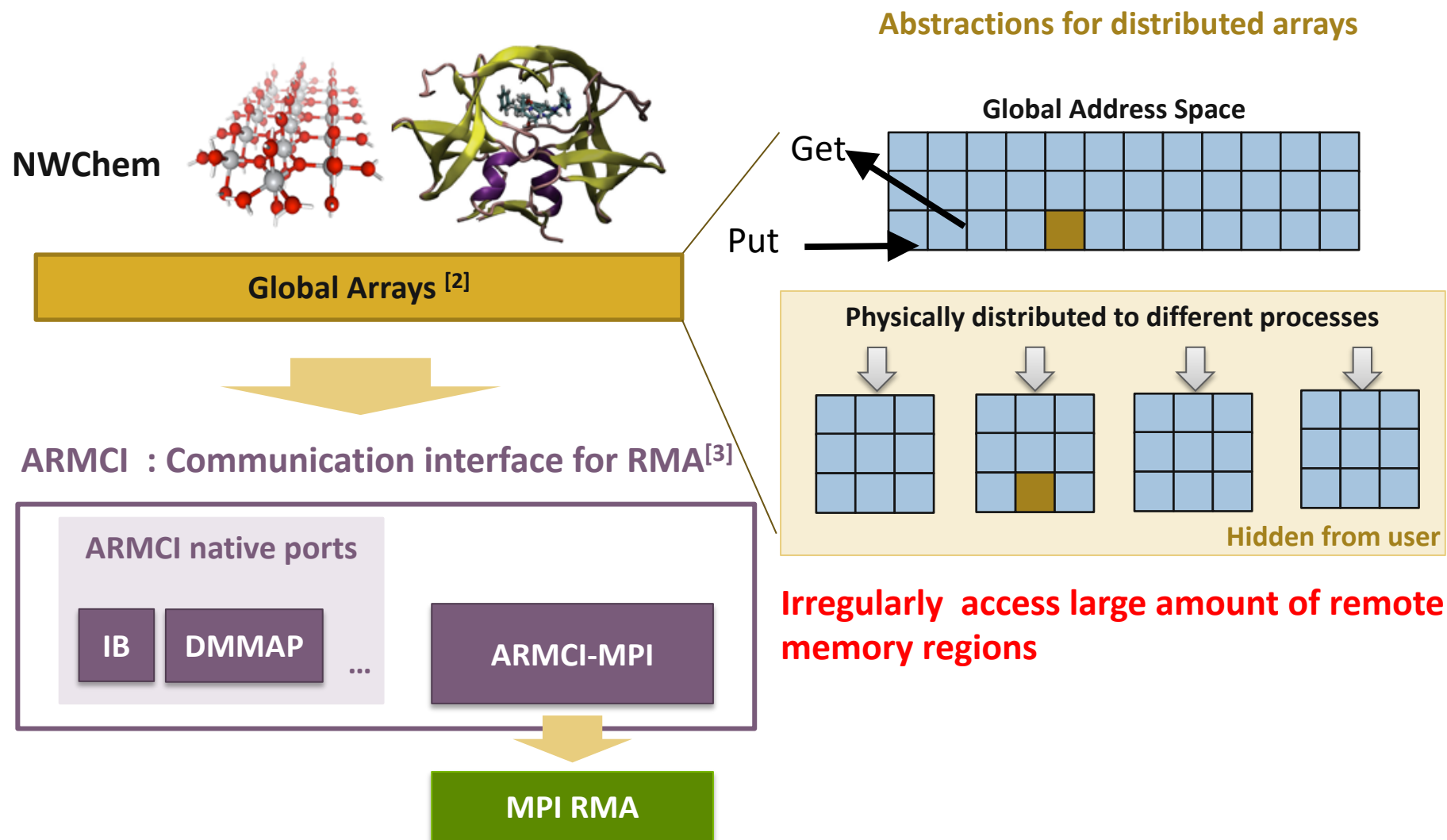
NWChem [1]

- High performance computational chemistry application suite
- Quantum level simulation of molecular systems
 - Very expensive in computation and data movement, so is used for small systems
 - Larger systems use molecular level simulations
- Composed of many simulation capabilities
 - Molecular Electronic Structure
 - Quantum Mechanics/Molecular Mechanics
 - Pseudo potential Plane-Wave Electronic Structure
 - Molecular Dynamics
- Very large code base
 - 4M LOC; Total investment of ~200M \$ to date



[1] M. Valiev, E.J. Bylaska, N. Govind, K. Kowalski, T.P. Straatsma, H.J.J. van Dam, D. Wang, J. Nieplocha, E. Apra, T.L. Windus, W.A. de Jong, "NWChem: a comprehensive and scalable open-source solution for large scale molecular simulations" Comput. Phys. Commun. 181, 1477 (2010)

NWChem Communication Runtime

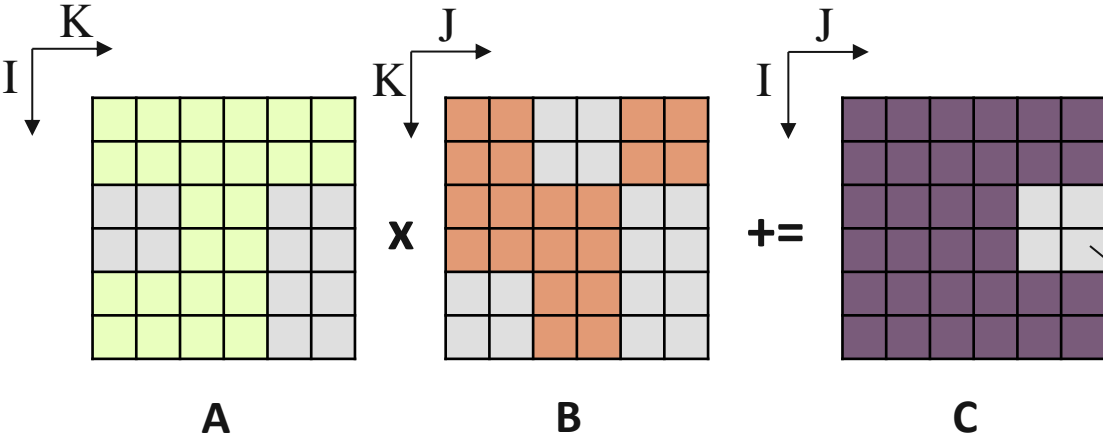


[2] <http://hpc.pnl.gov/globalarrays>

[3] <http://hpc.pnl.gov/armci>

Block Sparse Matrix Multiplication (BSPMM)

- Computing block sparse matrix multiplication: $C = \alpha A \times B + \beta C$



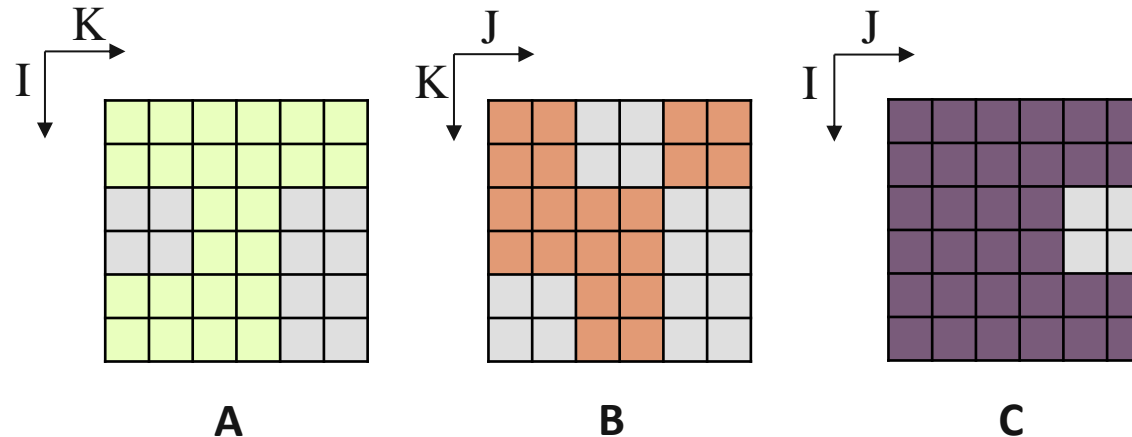
$A = \begin{bmatrix} a_{11} & \cdots & a_{1k} \\ \vdots & \ddots & \vdots \\ a_{i1} & \cdots & a_{ik} \end{bmatrix}; B = \begin{bmatrix} b_{11} & \cdots & b_{1j} \\ \vdots & \ddots & \vdots \\ b_{k1} & \cdots & b_{kj} \end{bmatrix}; C = \begin{bmatrix} c_{11} & \cdots & c_{1j} \\ \vdots & \ddots & \vdots \\ c_{i1} & \cdots & c_{ij} \end{bmatrix}$

$$c_{ij} = \alpha \sum_{k=1}^K a_{ik} b_{kj} + \beta c_{ij}, \quad \text{for } i = 1, \dots, I, \text{ and } j = 1, \dots, J.$$

Mock figure showing 2D DGEMM with block-sparse computations. In reality, NWChem uses 6D tensors.

The Global Data Structure

- Applications operate on large data sets that easily exceed the capacity of a single node
 - Resource sharing across nodes becomes necessary
 - “Allocate” matrices in the global address space that is physically distributed onto different nodes
- In the tutorial, the first process entirely allocates all three matrices for simplicity



The Parallel Approach

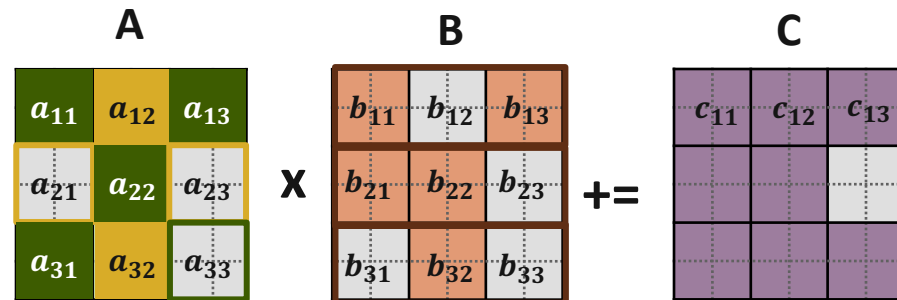
- Divide to independent subblocks and distribute over processes
- Compute DGEMM of subblocks locally
- For each c_{ij} , sum up the results of corresponding subblock multiplication

rank 0

```
c11+=a11*b11
Skip a11*(b12)
c13+=a11*b13
Skip a13*(b31)
c12+=a13*b32
Skip a13*(b33)
...
```

rank 1

```
c11+=a12*b21
c12+=a12*b22
Skip a12*(b23)
Skip (a21)
...
```

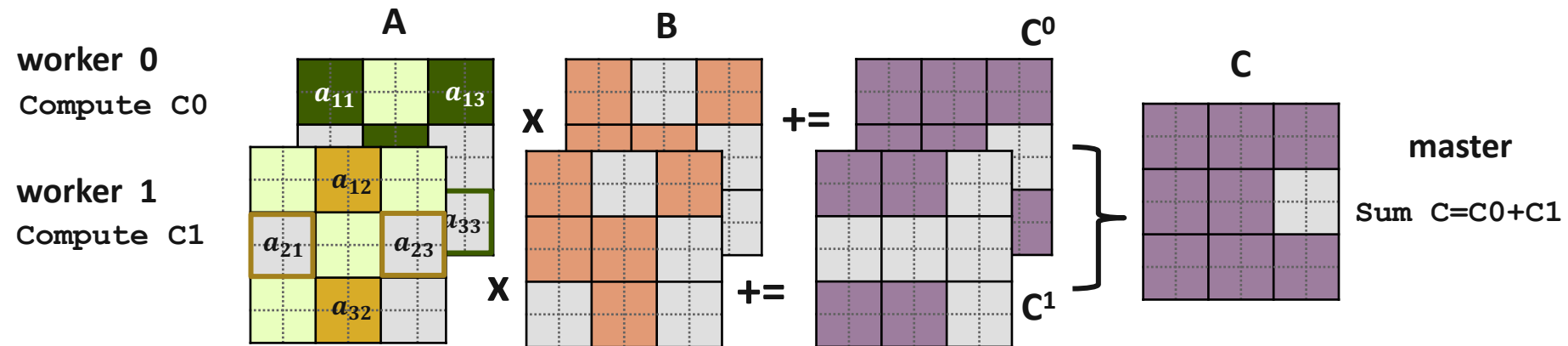


A-based parallelism

Set $\alpha=1$, $\beta=0$ for simplicity, compute $C = AB$

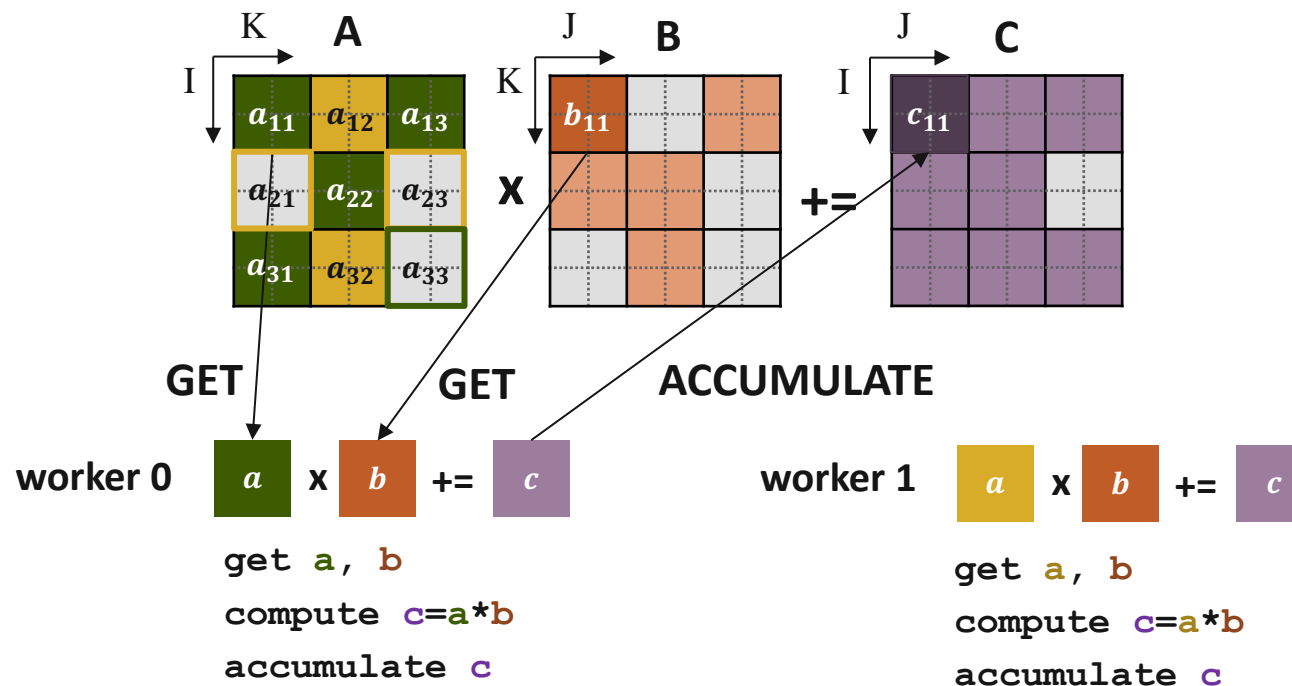
Example: BSPMM with full matrices

- *blocking_p2p/bspmm_simple.c*
- Simple bspmm code using blocking point-to-point operations
 - Rank 0 (master) initializes all matrices and sends the entire A and B matrices to other processes (workers)
 - Each worker computes different blocks and sends the local matrix C to master
 - Master sums up the received results



Example: BSPMM with RMA

- *rma/bspmm.c*
- Only synchronization from origin processes (workers), no synchronization from target processes (implicit master)
- Replace send/receive with RMA Get and Accumulate



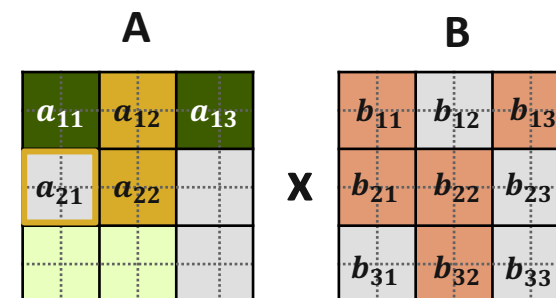
Exercise 5: BSPMM using global counter (atomics)

- In the basic RMA version of BSPMM, we statically assigned the ownership of blocks for each process
 - Load imbalance: e.g., one process may get more zero-blocks than others
- Let us implement a dynamic task secluding
 - Each process dynamically “queries” the next available block computation once finished the previous one (A-based parallelism)
 - Use atomic **Fetch_and_op(SUM)** with a “global counter”
- *Start from rma/bspmm.c*
- *Solution available in rma/bspmm_counter.c*

rank 0
FOP(counter) -> 0
compute **a11***b1j

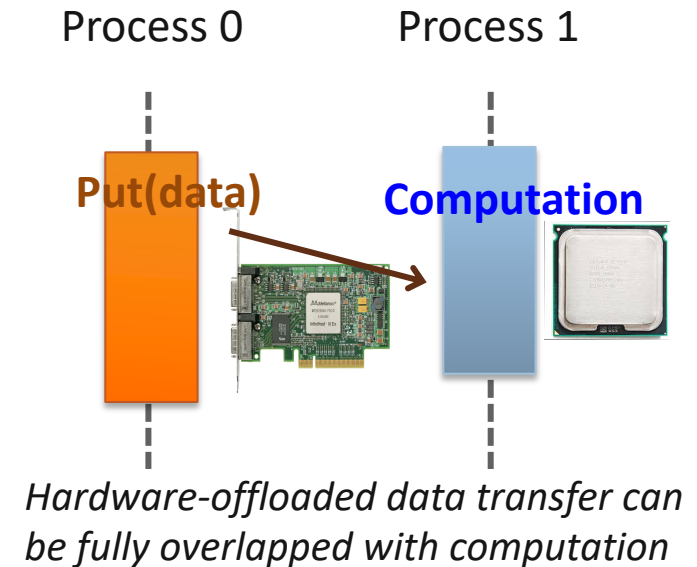
FOP(counter) -> 2
compute **a13***b3j
...

rank 1
FOP(counter) -> 1
compute **a12***b2j
FOP(counter) -> 3
Skip (a21)
FOP(counter) -> 4
compute **a22***b2j
...



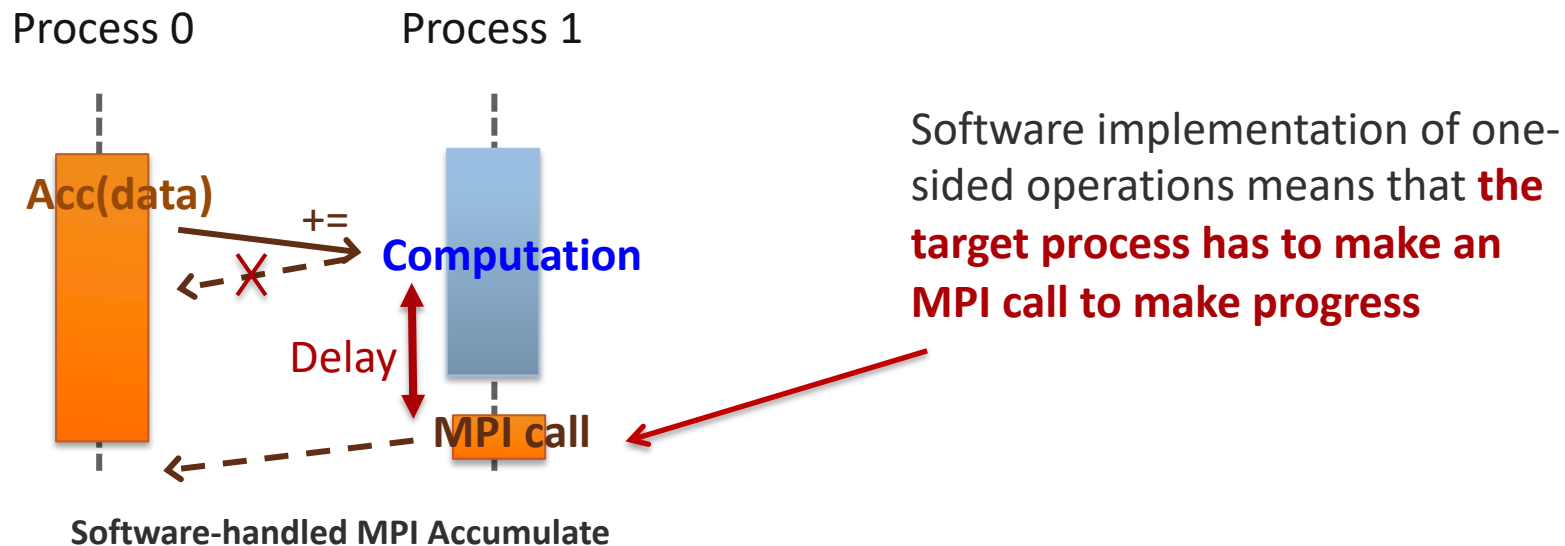
Hardware-Offloaded Communication

- Offloading data transfer to hardware is ideal for performance
 - Two-sided (e.g., SEND/RECEIVE):
 - Require **complex message matching** (rank + tag + comm), especially for wildcard receives (MPI_ANY_TAG|ANY_SOURCE)
 - Supported HW: Mellanox ConnectX-5 (support HW tag matching)
 - One-sided (e.g., PUT/GET/ACCUMUALTE):
 - **No matching** requirement, easier for hardware offloading
 - Natively supported on various RDMA networks such as Mellanox InfiniBand, Cray Aries, and Fujitsu Tofu



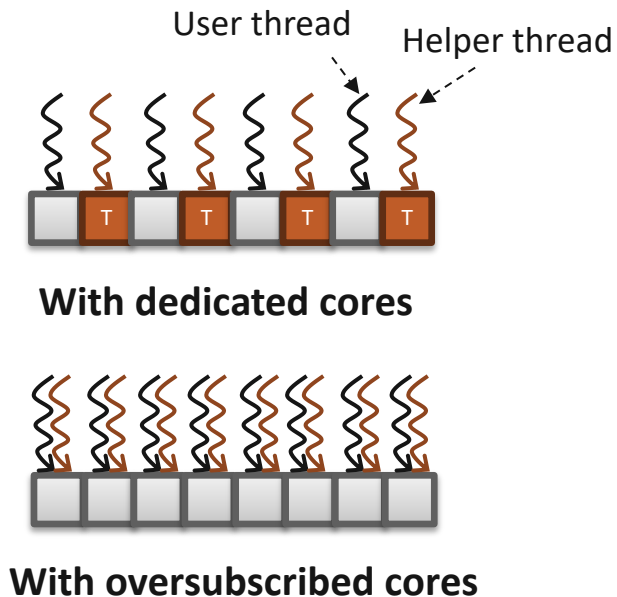
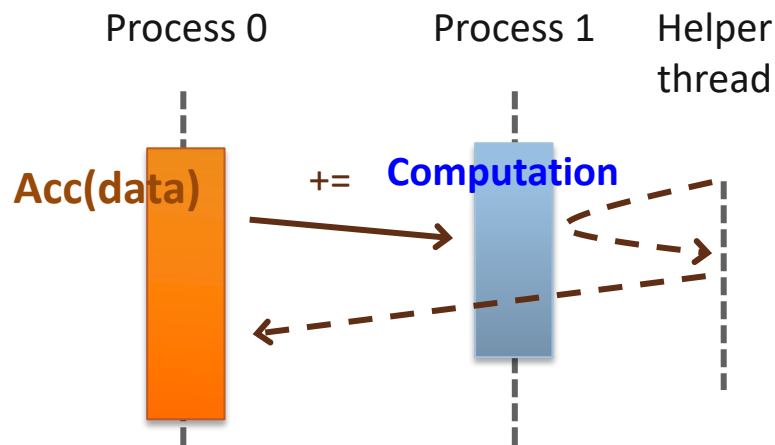
Asynchronous Execution of MPI RMA

- Asynchronous execution of RMA depends on the MPI implementation, which in turn depends on what the network hardware provides
- Most common situation on current network hardware:
 - Some operations are natively supported in hardware (e.g., contiguous PUT/GET)
 - Other operations need to be **emulated in software**



Possible Solution 1: Thread-based Progress (1/2)

- Every MPI process creates a **dedicated helper thread** at MPI_Init
 - Supported by most MPI implementations, but unlikely to be default
 - Might need to turn on some environment variable (check the documentation)
- The thread polls MPI progress for the process while the process is computing
- Dedicates some number of computing cores per process
- Multithreading safety overhead (i.e., MPI internal lock contention between threads, memory barriers)



Possible Solution 1: Thread-based Progress (2/2)

- Available in many mainstream MPI implementations
- Core binding is important! See vendor documentation

MPICH:

```
% export MPICH_CVAR_ASYNC_PROGRESS=1
```

Intel MPI:

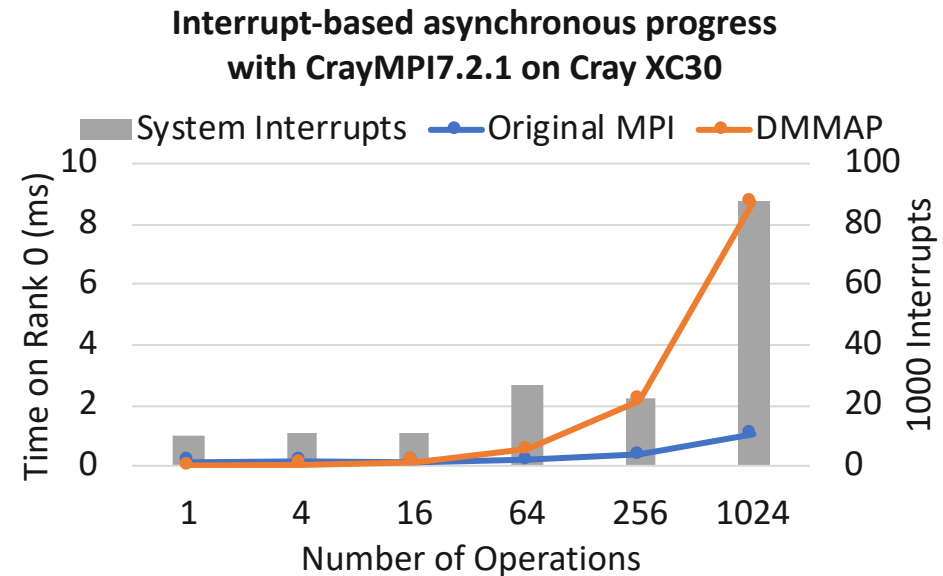
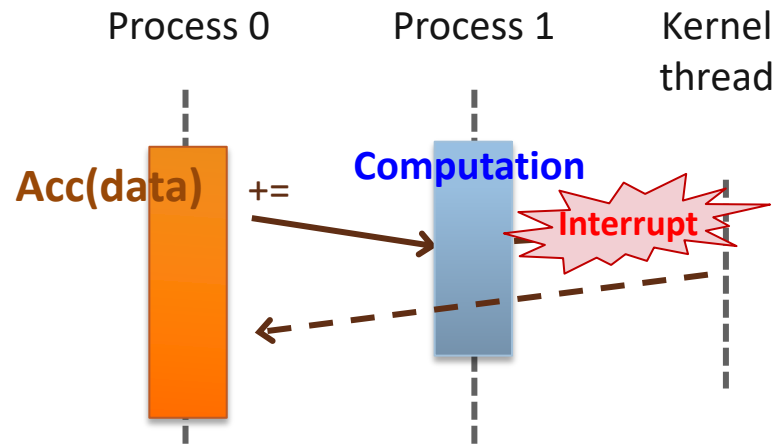
```
% export I_MPI_ASYNC_PROGRESS=1
```

Cray MPI:

```
% export MPICH_NEMESIS_ASYNC_PROGRESS=1  
% export MPICH_MAX_THREAD_SAFETY=multiple
```


Possible Solution 2: Interrupt-based Progress

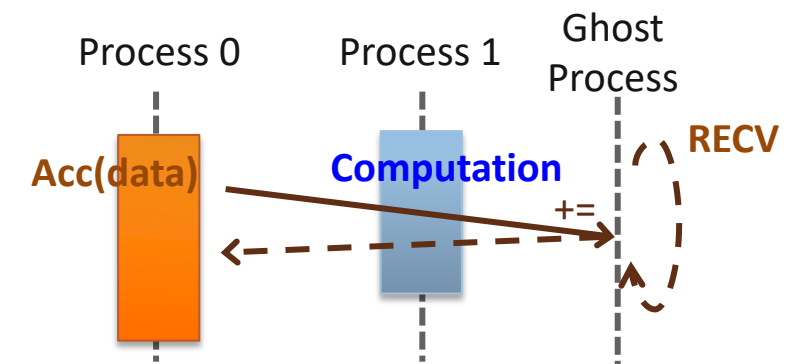
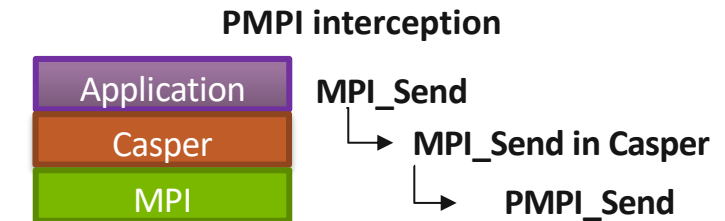
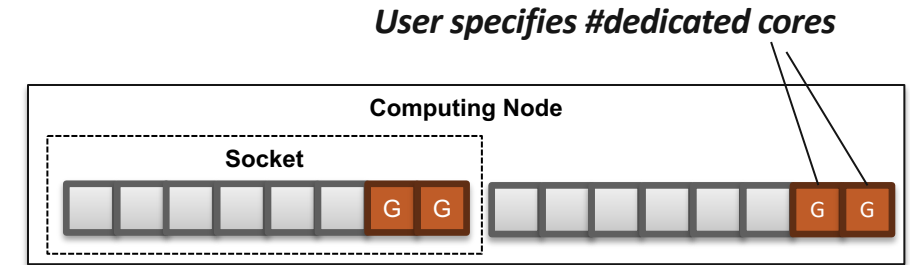
- Utilize **hardware interrupts** to awaken a kernel thread when new message arrives
- Examples: Cray MPI DMAPP mode (**deprecated from v7.6.0**), IBM MPI on Blue Gene/P
- Overhead of frequent interrupts, need special hardware support
 - Not a common model for most current networks



Possible Solution 3: Process-based Progress (1/4)

■ Casper

- Dedicating **arbitrary number of cores** (usually 1-4 per multicore node) to “**ghost processes**”
- **Portable PMPI profiling interface based** design allows integration with any MPI-3 implementations *
- Transparently replace MPI_COMM_WORLD by **COMM_USER_WORLD**
- **Shared memory mapping** between local user and ghost processes
- Redirect RMA operations to ghost process, thus ghost process ensures communication progress
- **Casper can redirect operations only for MPI_WIN_ALLOCATE windows**



* MPI-3 MPI_Win_allocate_shared function is used to enable shared memory mapping.

Possible Solution 3: Process-based Progress (2/4)

- Download Casper at <http://www.mcs.anl.gov/project/casper/downloads>
- How to use Casper?
 1. Load Casper between application and MPI library:
 - **Option 1:** dynamic load Casper at execution time

```
% mpicc -o app app.c
% export LD_PRELOAD=<casper_install_dir>/lib/libcasper.so
```

- **Option 2:** link with Casper (ensure link order: app.o -lcasper -lmpi)

```
% mpicc -o app app.c -lcasper
```

2. Set number of ghost processes (dedicated progress core)

```
% export CSP_NG=2
% mpiexec -np 48 -ppn 24 ./app
/* 48 processes are running on two nodes,
   * each node has 22 user processes and 2 ghost processes*/
```

Possible Solution 3: Process-based Progress (3/4)

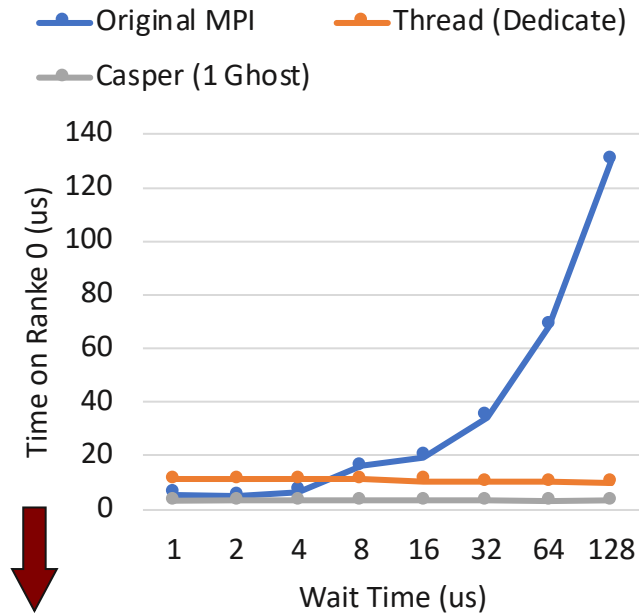
- Demonstrating communication progress

```
MPI_Win_allocate(sizeof(int), sizeof(int),
    MPI_INFO_NULL, MPI_COMM_WORLD, &a, &win);

if (rank == 0) {
    MPI_Win_lock(MPI_LOCK_SHARED, 1, 0, win);
    for (int x = 0; x < ITER; x++) {
        MPI_Accumulate(..., target = 1, ...);
        MPI_Win_flush(1, win);
    }
    MPI_Win_unlock(1, win);
    MPI_Send(..., dst = 1, ...);
} else {
    MPI_Irecv(..., src = 0, ..., &req);
    for (int x = 0; x < ITER; x++) {
        busy_wait(time);
        MPI_Test(&req, &flag, &stat);
    }
}

MPI_Win_free(&win);
```

MPI_Accumulate with asynchronous progress

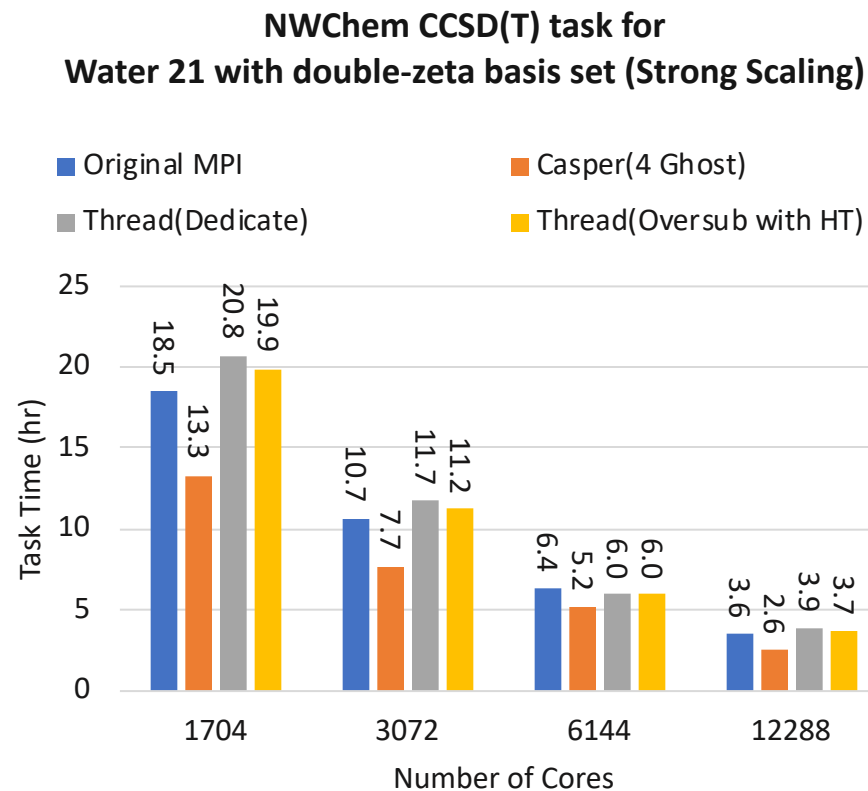
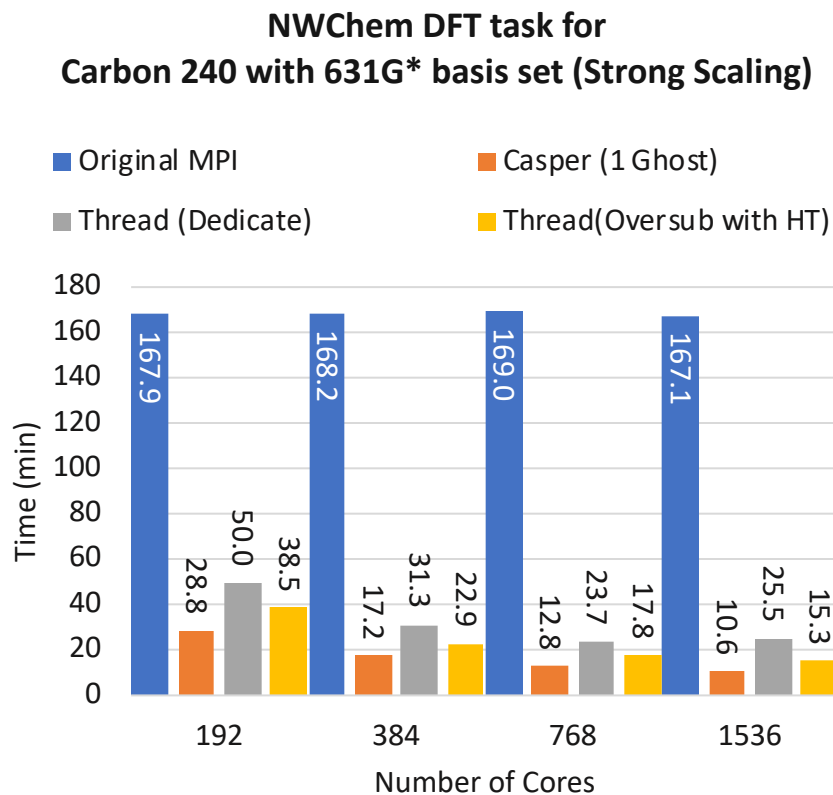


Lower is better

* Measured on NERSC Edison Cray XC30 with CrayMPI 6.3.1.

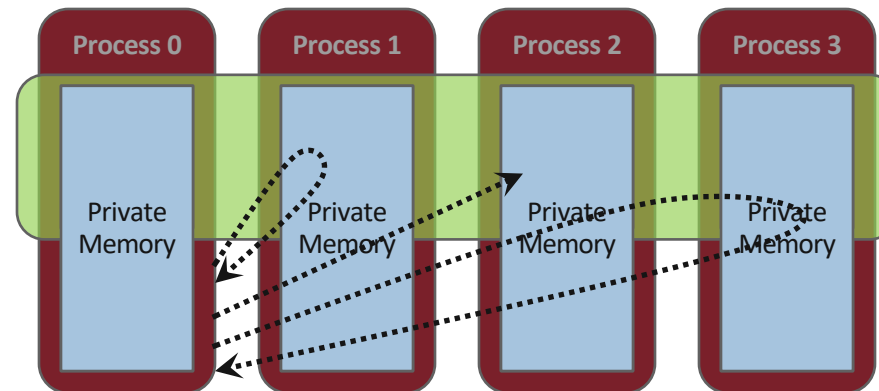
Possible Solution 3: Process-based Progress (4/4)

- Performance improvement of NWChem with RMA asynchronous progress



Section Summary

- MPI one-sided communication is associated with **windows**
- **Operations** include basic **PUT**, **GET**, and **Atomic** operations
- **Synchronization** modes
 - Active-target (similar to two-sided) : FENCE, PSCW
 - Passive-target: LOCK-UNLOCK, FLUSH, FLUSH_LOCAL...
- Enable **asynchronous progress** for performance

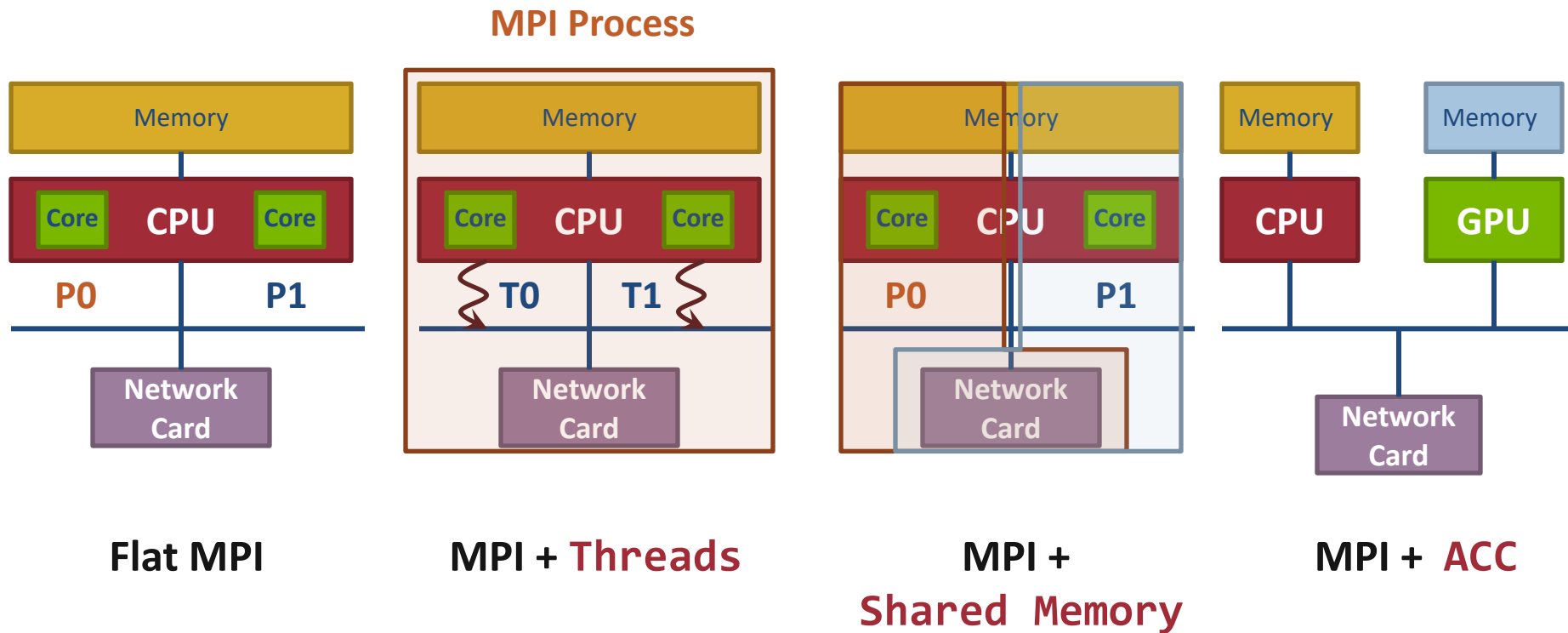


MPI Hybrid Programming: Threads



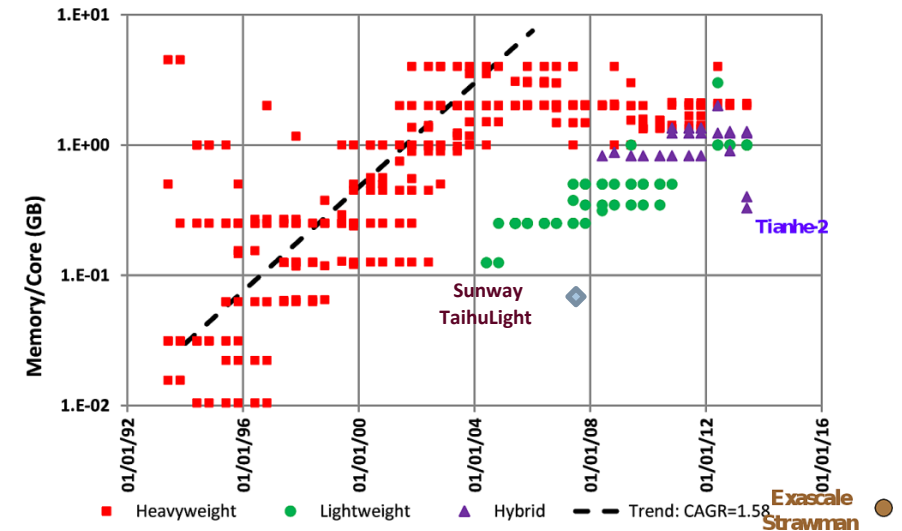
Hybrid MPI + X : Most Popular Forms

MPI + X



Why Hybrid MPI+X? Towards Strong Scaling (1/3)

- Strong scaling applications is increasing in importance
 - Hardware limitations: not all resources scale at the same rate as cores (e.g., memory capacity, network resources)
 - Desire to solve the same problem faster on a bigger machine
 - Nek5000, HACC, LAMMPS
- Strong scaling pure MPI applications is getting harder
 - On-node communication is costly compared to load/stores
 - $O(Px)$ communication patterns (e.g., All-to-all) costly

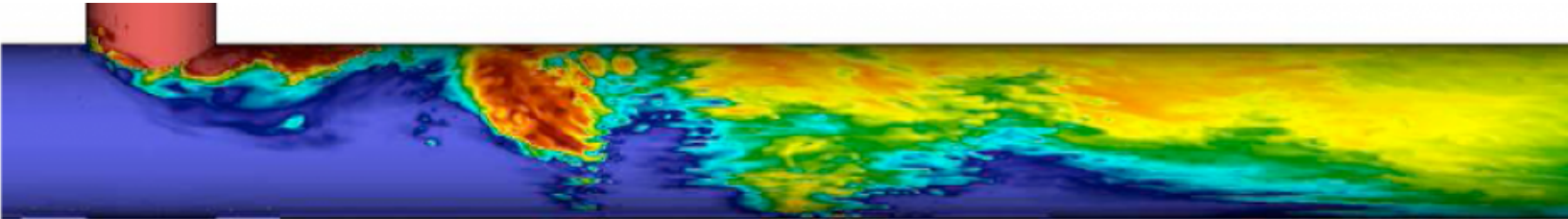


Evolution of the memory capacity per core in the Top500 list (Peter Kogge. PIM & memory: The need for a revolution in architecture.)

Why Hybrid MPI+X? Towards Strong Scaling (2/3)

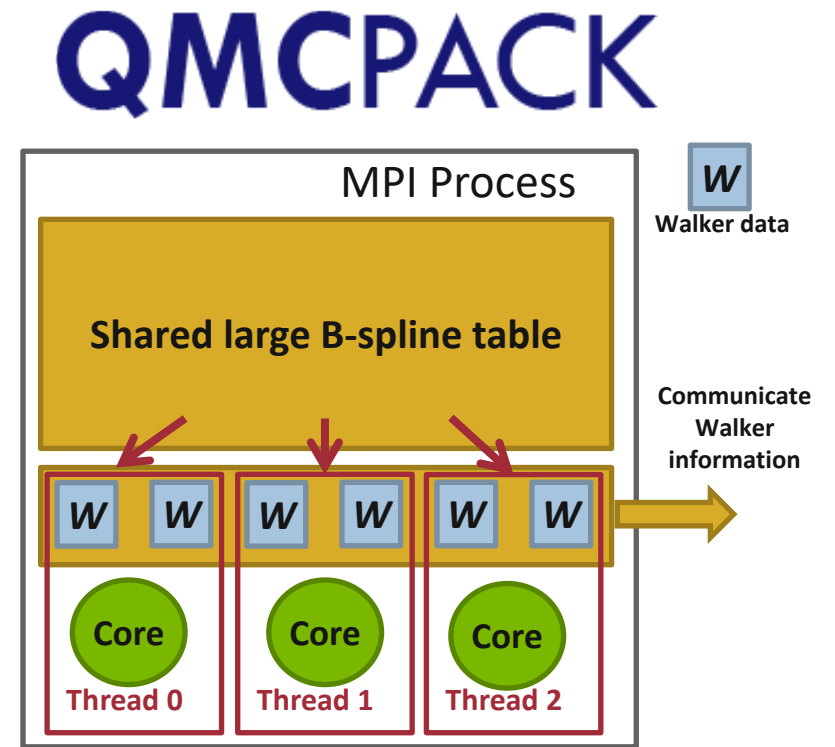
- MPI+X benefits ($X = \{\text{threads, MPI shared-memory, etc.}\}$)
 - Less memory hungry (MPI runtime consumption, $O(P)$ data structures, etc.)
 - Load/stores to access memory instead of message passing
 - P is reduced by constant C (#cores/process) for $O(Px)$ communication patterns
- Example 1: the Nek5000 team is working at the strong scaling limit

Nek5000



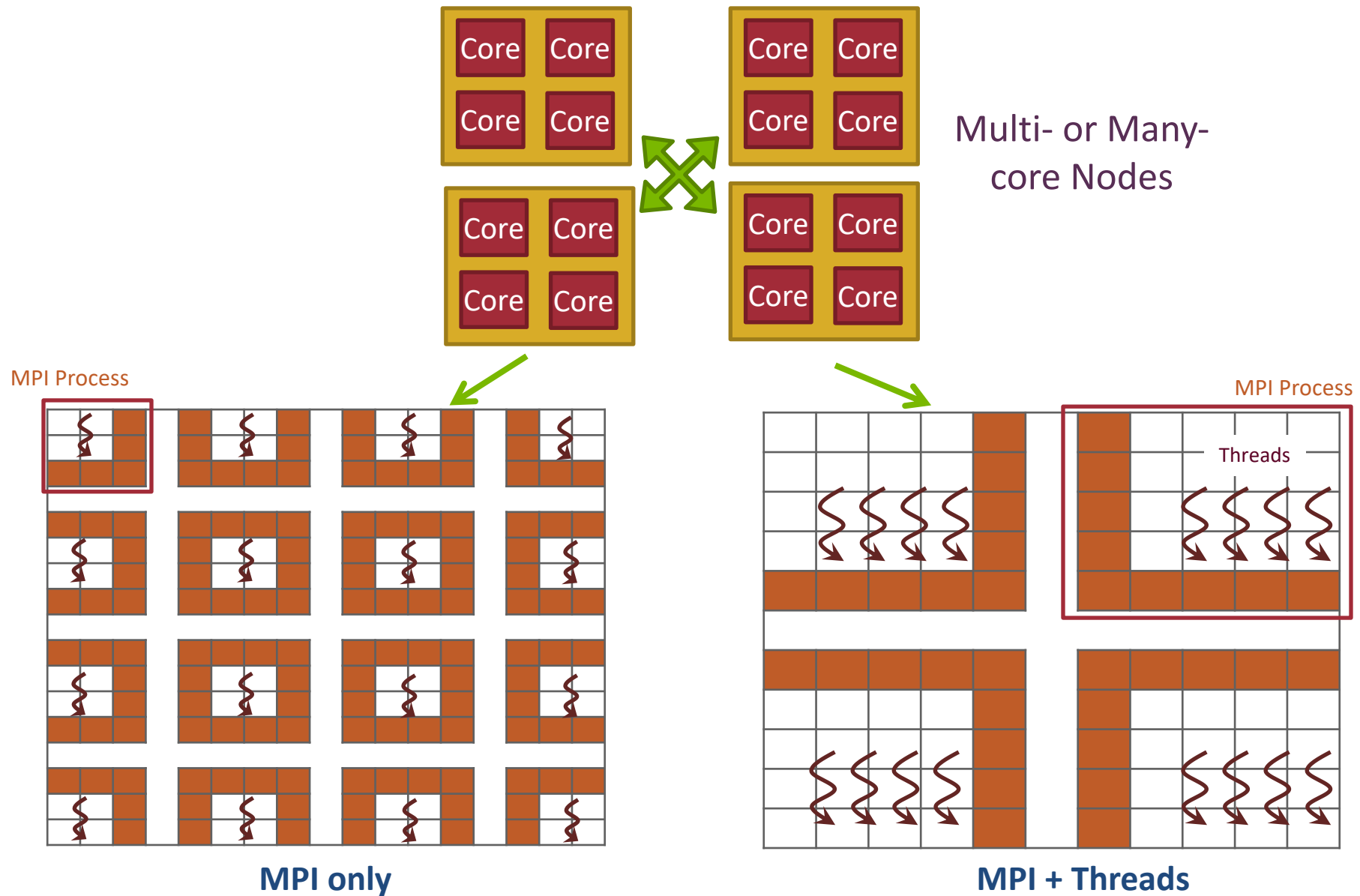
Why Hybrid MPI+X? Towards Strong Scaling (3/3)

- Example 2: Quantum Monte Carlo Simulation (QMCPACK)
 - Size of the physical system to simulate is bound by memory capacity [1]
 - Memory space dominated by large interpolation tables (typically several Giga Bytes of storage)
 - Threads are used to share those tables
 - Memory for communication buffers must be kept low to be allow simulation of larger and highly detailed simulations.



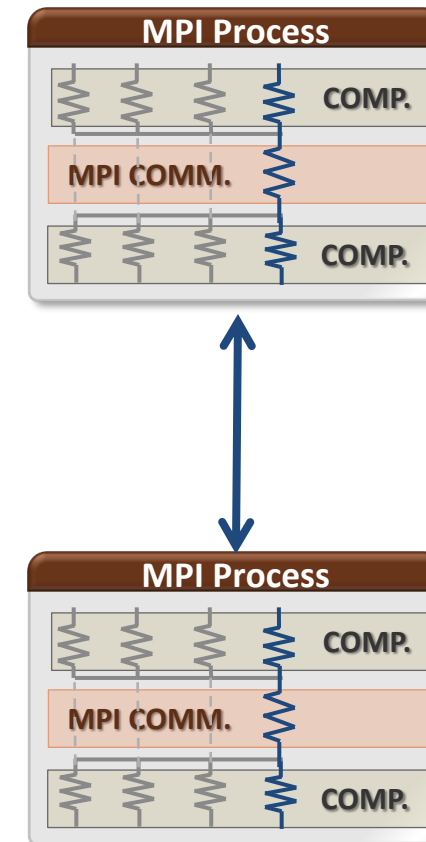
[1] Kim, Jeongnim, et al. "Hybrid algorithms in quantum Monte Carlo." Journal of Physics, 2012.

MPI + Threads: How To? (1/3)



MPI + Threads: How To? (2/3)

- MPI describes parallelism between *processes* (with separate address spaces)
- *Thread* parallelism provides a shared-memory model within a process
- OpenMP and Pthreads are common models
 - OpenMP provides convenient features for loop-level parallelism. Threads are created and managed by the compiler, based on user directives.
 - Pthreads provide more complex and dynamic approaches. Threads are created and managed explicitly by the user.



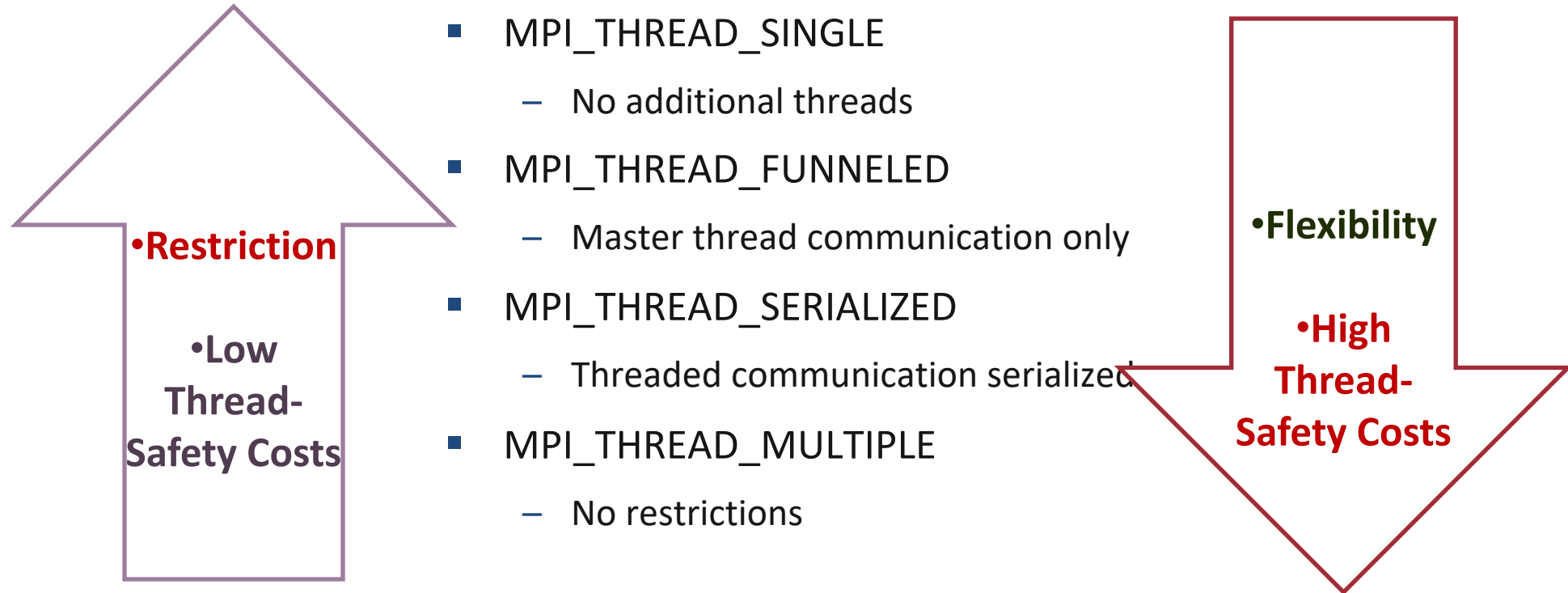
MPI + Threads: How To? (3/3)

MPI + Threads



Interoperability

Interoperation or thread levels:



MPI's Four Levels of Thread Safety

- MPI defines four levels of thread safety -- these are commitments the application makes to the MPI
- Thread levels are in increasing order
 - If an application works in FUNNELED mode, it can work in SERIALIZED
- MPI defines an alternative to MPI_Init
 - **MPI_Init_thread**(requested, provided): *Application specifies level it needs; MPI implementation returns level it supports*

MPI_THREAD_SINGLE

- There are no additional user threads in the system
 - E.g., there are no OpenMP parallel regions

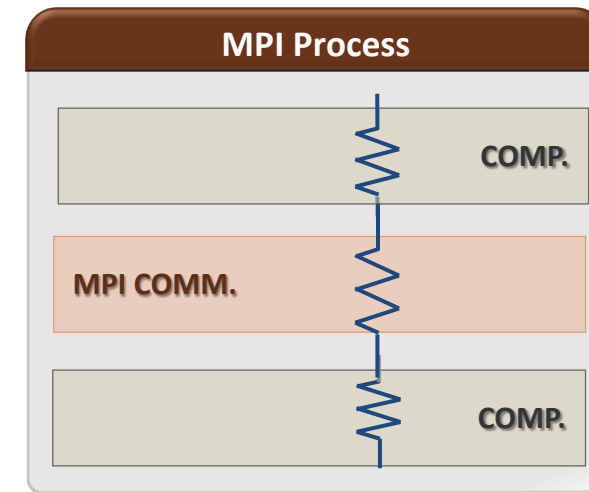
```
int buf[100];
int main(int argc, char ** argv)
{
    MPI_Init(&argc, &argv);
    MPI_Comm_rank(MPI_COMM_WORLD, &rank);

    for (i = 0; i < 100; i++)
        compute(buf[i]);

    /* Do MPI stuff */

    MPI_Finalize();

    return 0;
}
```



MPI_THREAD_FUNNELED

- All MPI calls are made by the **master** thread
 - Outside the OpenMP parallel regions
 - In OpenMP master regions

```
int buf[100];
int main(int argc, char ** argv)
{
    int provided;

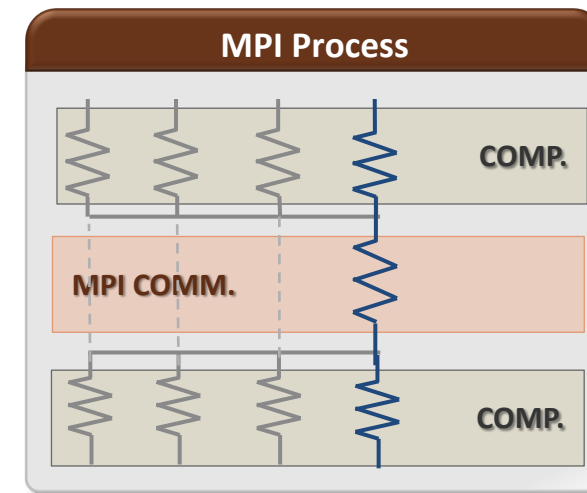
    MPI_Init_thread(&argc, &argv,
        MPI_THREAD_FUNNELED, &provided);
    if (provided < MPI_THREAD_FUNNELED)
        MPI_Abort(MPI_COMM_WORLD, 1);

    for (i = 0; i < 100; i++)
        pthread_create(..., func, (void*)i);
    for (i = 0; i < 100; i++)
        pthread_join();

    /* Do MPI stuff */

    MPI_Finalize();
    return 0;
}
```

```
void* func(void* arg) {
    int i = (int)arg;
    compute(buf[i]);
}
```



MPI_THREAD_SERIALIZED

- Only **one** thread can make MPI calls at a time
 - Protected by OpenMP critical regions

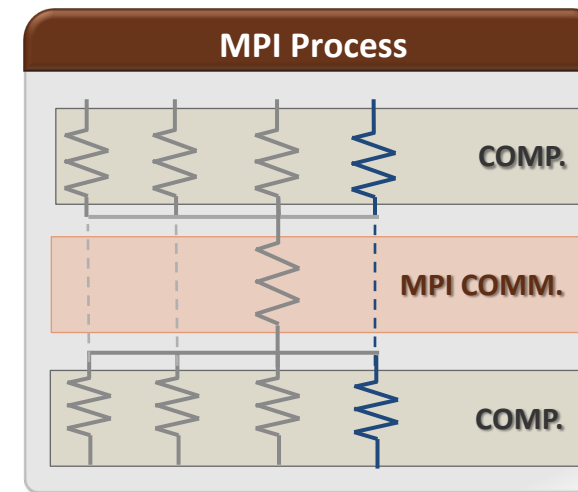
```
int buf[100];
int main(int argc, char ** argv)
{
    int provided;
    pthread_mutex_t mutex;

    MPI_Init_thread(&argc, &argv,
        MPI_THREAD_SERIALIZED, &provided);
    if (provided < MPI_THREAD_SERIALIZED)
        MPI_Abort(MPI_COMM_WORLD, 1);

    for (i = 0; i < 100; i++)
        pthread_create(..., func, (void*)i);
    for (i = 0; i < 100; i++)
        pthread_join();

    MPI_Finalize();
    return 0;
}
```

```
void* func(void* arg) {
    int i = (int)arg;
    compute(buf[i]);
    pthread_mutex_lock(&mutex);
    /* Do MPI stuff */
    pthread_mutex_unlock(&mutex);
}
```



MPI_THREAD_MULTIPLE

- **Any** thread can make MPI calls any time (restrictions apply)

```
int buf[100];
int main(int argc, char ** argv)
{
    int provided;

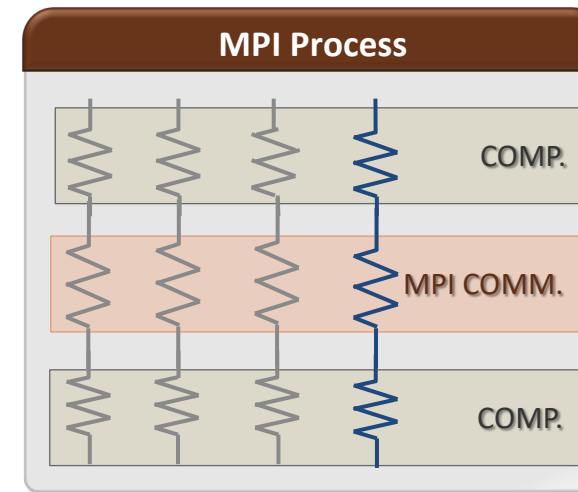
    MPI_Init_thread(&argc, &argv,
        MPI_THREAD_MULTIPLE, &provided);
    if (provided < MPI_THREAD_MULTIPLE)
        MPI_Abort(MPI_COMM_WORLD, 1);

    for (i = 0; i < 100; i++)
        pthread_create(..., func, (void*)i);
    for (i = 0; i < 100; i++)
        pthread_join();

    MPI_Finalize();
    return 0;
}
```

```
void* func(void* arg) {
    int i = (int)arg;
    compute(buf[i]);

    /* Do MPI stuff */
}
```



Threads and MPI

- An implementation is not required to support levels higher than MPI_THREAD_SINGLE; that is, an implementation is not required to be thread safe
- A fully thread-safe implementation will support MPI_THREAD_MULTIPLE
- A program that calls MPI_Init (instead of MPI_Init_thread) should assume that only MPI_THREAD_SINGLE is supported
 - MPI Standard *mandates* MPI_THREAD_SINGLE for MPI_Init
- *A threaded MPI program that does not call MPI_Init_thread is an incorrect program (common user error we see)*

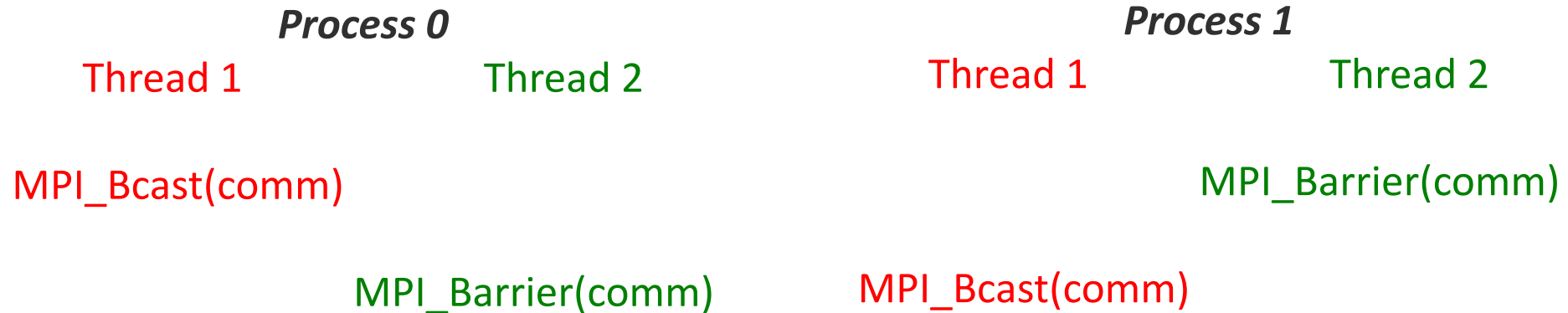
MPI Semantics and `MPI_THREAD_MULTIPLE`

- **Ordering:** When multiple threads make MPI calls concurrently, the outcome will be as if the calls executed sequentially in some (any) order
 - Ordering is maintained within each thread
 - User must ensure that collective operations on the same communicator, window, or file handle are correctly ordered among threads
 - E.g., cannot call a broadcast on one thread and a reduce on another thread on the same communicator
 - It is the user's responsibility to prevent races when threads in the same application post conflicting MPI calls
 - E.g., accessing an info object from one thread and freeing it from another thread
- **Progress:** Blocking MPI calls will block only the calling thread and will not prevent other threads from running or executing MPI functions

Ordering in MPI_THREAD_MULTIPLE: Incorrect Example with Collectives

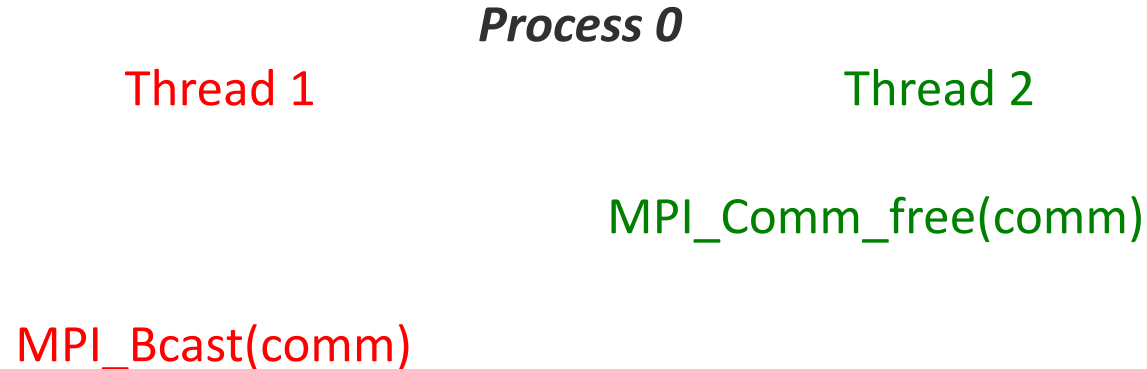
	<i>Process 0</i>	<i>Process 1</i>
<i>Thread 0</i>	MPI_Bcast(comm)	MPI_Bcast(comm)
<i>Thread 1</i>	MPI_Barrier(comm)	MPI_Barrier(comm)

Ordering in MPI_THREAD_MULTIPLE: Incorrect Example with Collectives



- P0 and P1 can have different orderings of Bcast and Barrier
- Here the user must use some kind of synchronization to ensure that either thread 1 or thread 2 gets scheduled first on both processes
- Otherwise a broadcast may get matched with a barrier on the same communicator, which is not allowed in MPI

Ordering in MPI_THREAD_MULTIPLE: Incorrect Example with Object Management



- The user has to make sure that one thread is not using an object while another thread is freeing it
 - This is essentially an ordering issue; the object might get freed before it is used

Blocking Calls in MPI_THREAD_MULTIPLE: Correct Example

	<i>Process 0</i>	<i>Process 1</i>
Thread 1	MPI_Recv(src=1)	MPI_Recv(src=0)
Thread 2	MPI_Send(dst=1)	MPI_Send(dst=0)

- An implementation must ensure that the above example never deadlocks for any ordering of thread execution
- That means the implementation cannot simply acquire a thread lock and block within an MPI function. It must release the lock to allow other threads to make progress.

The Current Situation

- All MPI implementations support `MPI_THREAD_SINGLE`
- They probably support `MPI_THREAD_FUNNELED` even if they don't admit it.
 - Does require thread-safety for some system routines (e.g. `malloc`)
 - On most systems `-pthread` will guarantee it (OpenMP implies `-pthread`)
- Many (but not all) implementations support `THREAD_MULTIPLE`
 - Hard to implement efficiently though (thread synchronization issues)
- Bulk-synchronous OpenMP programs (loops parallelized with OpenMP, communication between loops) only need `FUNNELED`
 - So don't need "thread-safe" MPI for many hybrid programs
 - But watch out for Amdahl's Law!

Hybrid Programming: Correctness Requirements

- Hybrid programming with MPI+threads does not do much to reduce the complexity of thread programming
 - Your application still has to be a correct multi-threaded application
 - On top of that, you also need to make sure you are correctly following MPI semantics
- Many commercial debuggers offer support for debugging hybrid MPI+threads applications (mostly for MPI+Pthreads and MPI+OpenMP)

An Example we encountered

- We received a bug report about a very simple multithreaded MPI program that hangs
- Run with 2 processes
- Each process has 2 threads
- Both threads communicate with threads on the other process as shown in the next slide
- We spent several hours trying to debug MPICH before discovering that the bug is actually in the user's program 😞

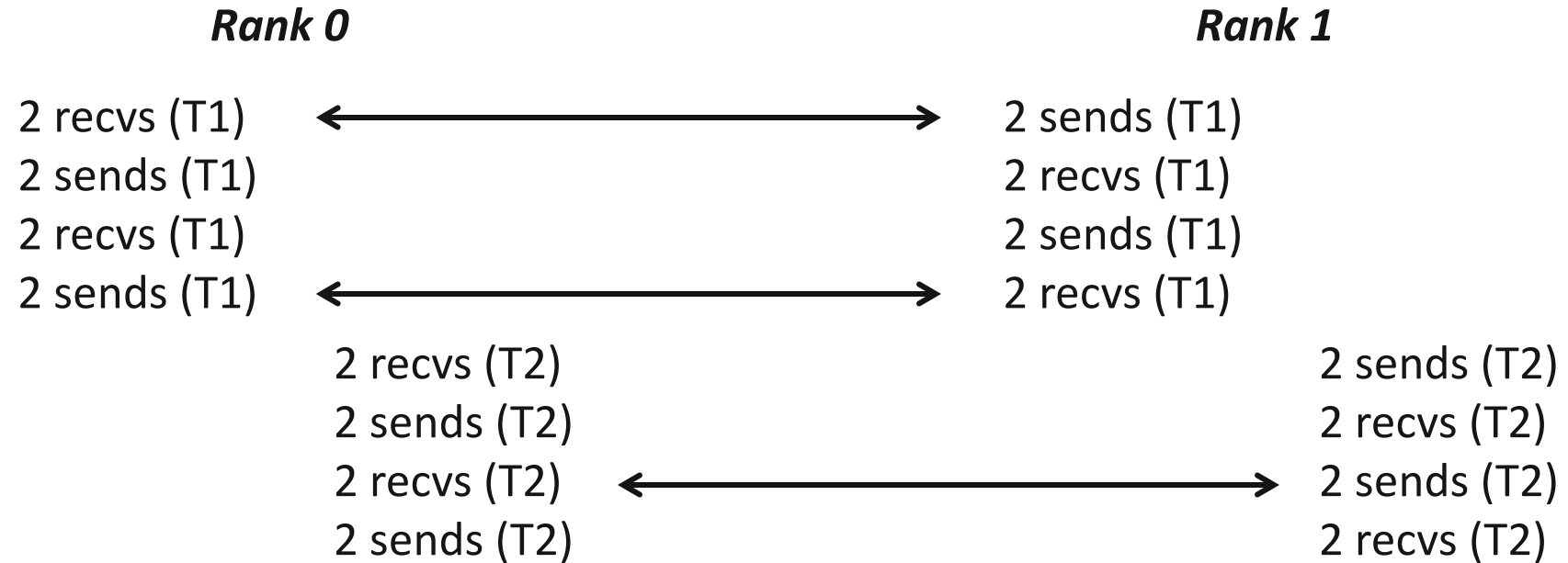
2 Processes, 2 Threads (each thread executes this code)

```
if (rank == 1) {
    MPI_Send(NULL, 0, MPI_CHAR, 0, 0, MPI_COMM_WORLD);
    MPI_Send(NULL, 0, MPI_CHAR, 0, 0, MPI_COMM_WORLD);
    MPI_Recv(NULL, 0, MPI_CHAR, 0, 0, MPI_COMM_WORLD, &stat);
    MPI_Recv(NULL, 0, MPI_CHAR, 0, 0, MPI_COMM_WORLD, &stat);

    MPI_Send(NULL, 0, MPI_CHAR, 0, 0, MPI_COMM_WORLD);
    MPI_Send(NULL, 0, MPI_CHAR, 0, 0, MPI_COMM_WORLD);
    MPI_Recv(NULL, 0, MPI_CHAR, 0, 0, MPI_COMM_WORLD, &stat);
    MPI_Recv(NULL, 0, MPI_CHAR, 0, 0, MPI_COMM_WORLD, &stat);
} else { /* rank == 0 */
    MPI_Recv(NULL, 0, MPI_CHAR, 1, 0, MPI_COMM_WORLD, &stat);
    MPI_Recv(NULL, 0, MPI_CHAR, 1, 0, MPI_COMM_WORLD, &stat);
    MPI_Send(NULL, 0, MPI_CHAR, 1, 0, MPI_COMM_WORLD);
    MPI_Send(NULL, 0, MPI_CHAR, 1, 0, MPI_COMM_WORLD);

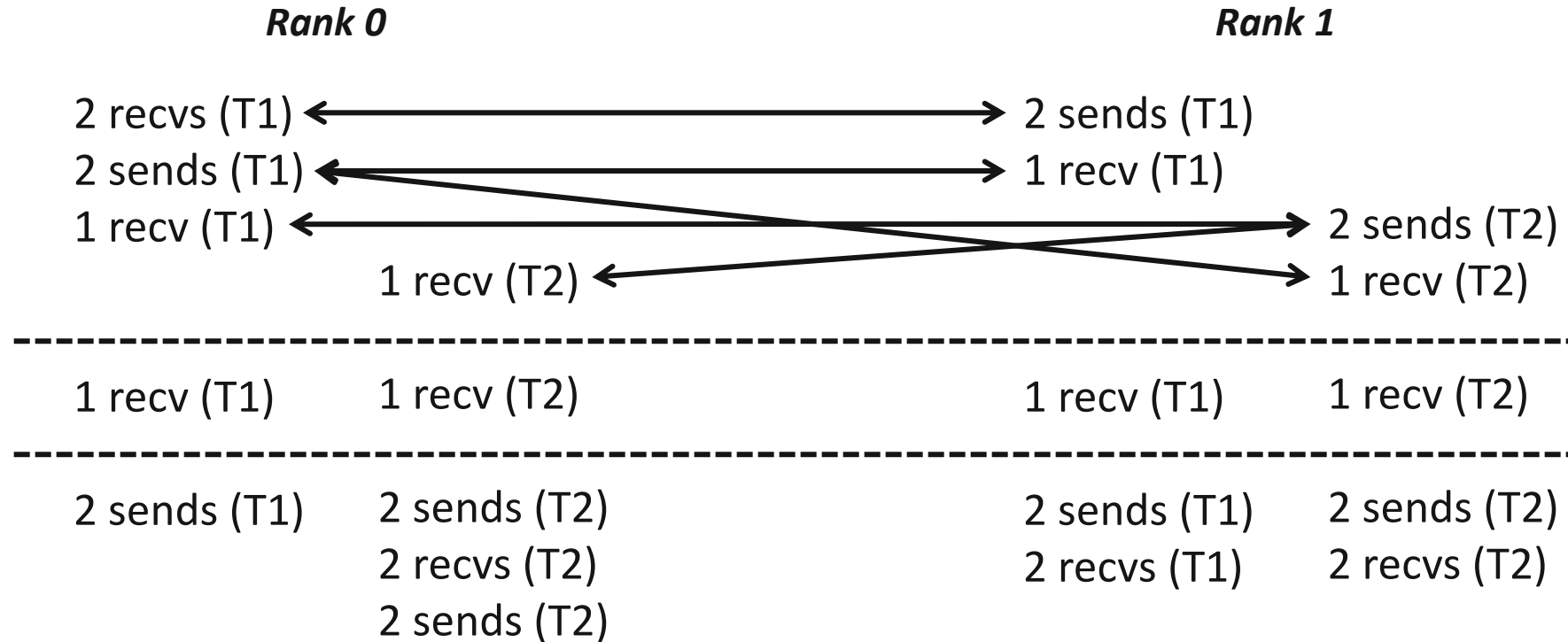
    MPI_Recv(NULL, 0, MPI_CHAR, 1, 0, MPI_COMM_WORLD, &stat);
    MPI_Recv(NULL, 0, MPI_CHAR, 1, 0, MPI_COMM_WORLD, &stat);
    MPI_Send(NULL, 0, MPI_CHAR, 1, 0, MPI_COMM_WORLD);
    MPI_Send(NULL, 0, MPI_CHAR, 1, 0, MPI_COMM_WORLD);
}
```

Intended Ordering of Operations



- Every send matches a receive on the other rank

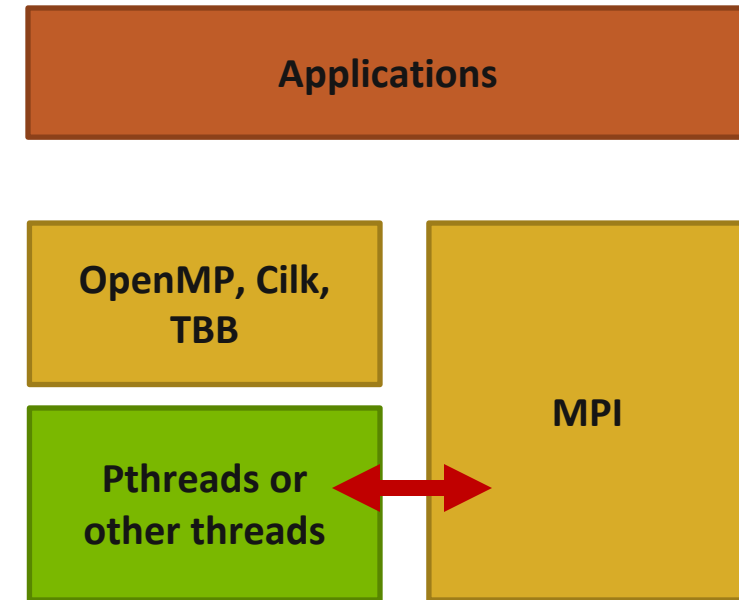
Possible Ordering of Operations in Practice



- Because the MPI operations can be issued in an arbitrary order across threads, all threads could block in a RECV call

MPI+OpenMP correctness semantics

- MPI only specifies interoperability with threads, not with OpenMP (or any other high-level programming model using threads)
 - OpenMP iterations need to be carefully mapped to which thread executes them (some schedules in OpenMP make this harder)
- For OpenMP tasks, the general model to use is that an OpenMP thread can execute one or more OpenMP tasks
 - An MPI blocking call should be assumed to block the entire OpenMP thread, so other tasks might not get executed



OpenMP threads: MPI blocking Calls (1/2)

```
int main(int argc, char ** argv)
{
    MPI_Init_thread(NULL, NULL, MPI_THREAD_MULTIPLE, &provided);

    #pragma omp parallel for
    for (i = 0; i < 100; i++) {
        if (i % 2 == 0)
            MPI_Send(.., to_myself, ..);
        else
            MPI_Recv(.., from_myself, ..);
    }

    MPI_Finalize();

    return 0;
}
```

Iteration to OpenMP thread mapping needs to explicitly be handled by the user; otherwise, OpenMP threads might all issue the same operation and deadlock

OpenMP threads: MPI blocking Calls (2/2)

```
int main(int argc, char ** argv)
{
    MPI_Init_thread(NULL, NULL, MPI_THREAD_MULTIPLE, &provided);

#pragma omp parallel
{
    assert(omp_get_num_threads() > 1)
#pragma omp for schedule(static, 1)
    for (i = 0; i < 100; i++) {
        if (i % 2 == 0)
            MPI_Send(.., to_myself, ..);
        else
            MPI_Recv(.., from_myself, ..);
    }
}

    MPI_Finalize();

    return 0;
}
```

Either explicit/careful mapping of iterations to threads, or using nonblocking versions of send/recv would solve this problem

OpenMP tasks: MPI blocking Calls (1/5)

```
int main(int argc, char ** argv)
{
    MPI_Init_thread(NULL, NULL, MPI_THREAD_MULTIPLE, &provided);

#pragma omp parallel
{
    #pragma omp for
    for (i = 0; i < 100; i++) {
        #pragma omp task
        {
            if (i % 2 == 0)
                MPI_Send(..., to_myself, ..);
            else
                MPI_Recv(..., from_myself, ..);
        }
    }
}

MPI_Finalize();
return 0;
}
```

This can lead to deadlocks. No ordering or progress guarantees in OpenMP task scheduling should be assumed; a blocked task blocks it's thread and tasks can be executed in any order.

OpenMP tasks: MPI blocking Calls (2/5)

```
int main(int argc, char ** argv)
{
    MPI_Init_thread(NULL, NULL, MPI_THREAD_MULTIPLE, &provided);

#pragma omp parallel
{
    #pragma omp taskloop
    for (i = 0; i < 100; i++) {
        if (i % 2 == 0)
            MPI_Send(.., to_myself, ..);
        else
            MPI_Recv(.., from_myself, ..)
    }
}

    MPI_Finalize();
    return 0;
}
```

Same problem as before.

OpenMP tasks: MPI blocking Calls (3/5)

```
int main(int argc, char ** argv)
{
    MPI_Init_thread(NULL, NULL, MPI_THREAD_MULTIPLE, &provided);

#pragma omp parallel
{
    #pragma omp taskloop
    for (i = 0; i < 100; i++) {
        MPI_Request req;
        if (i % 2 == 0)
            MPI_Isend(.., to_myself, .., &req);
        else
            MPI_Irecv(.., from_myself, .., &req);
        MPI_Wait(&req, ..);
    }
}

MPI_Finalize();
return 0;
}
```

Using nonblocking operations but with MPI_Wait inside the task region does not solve the problem

OpenMP tasks: MPI blocking Calls (4/5)

```
int main(int argc, char ** argv)
{
    MPI_Init_thread(NULL, NULL, MPI_THREAD_MULTIPLE, &provided);

#pragma omp parallel
{
    #pragma omp taskloop
    for (i = 0; i < 100; i++) {
        MPI_Request req; int done = 0;
        if (i % 2 == 0)
            MPI_Isend(.., to_myself, .., &req);
        else
            MPI_Irecv(.., from_myself, .., &req);
        While (!done) {
            #pragma omp taskyield
            MPI_Test(&req, &done, ..);
        }
    }
}

MPI_Finalize();
return 0;
}
```

Still incorrect; taskyield does not guarantee a task switch

OpenMP tasks: MPI blocking Calls (5/5)

```
int main(int argc, char ** argv)
{
    MPI_Init_thread(NULL, NULL, MPI_THREAD_MULTIPLE, &provided);
    MPI_Request req[100];

#pragma omp parallel
{
    #pragma omp taskloop
    for (i = 0; i < 100; i++) {
        if (i % 2 == 0)
            MPI_Isend(.., to_myself, .., &req[i]);
        else
            MPI_Irecv(.., from_myself, .., &req[i]);
    }
}

MPI_Waitall(100, req, ..);
MPI_Finalize();
return 0;
}
```

Correct example. Each task is nonblocking.

Ordering in MPI_THREAD_MULTIPLE: Incorrect Example with RMA

```
int main(int argc, char ** argv)
{
    /* Initialize MPI and RMA window */

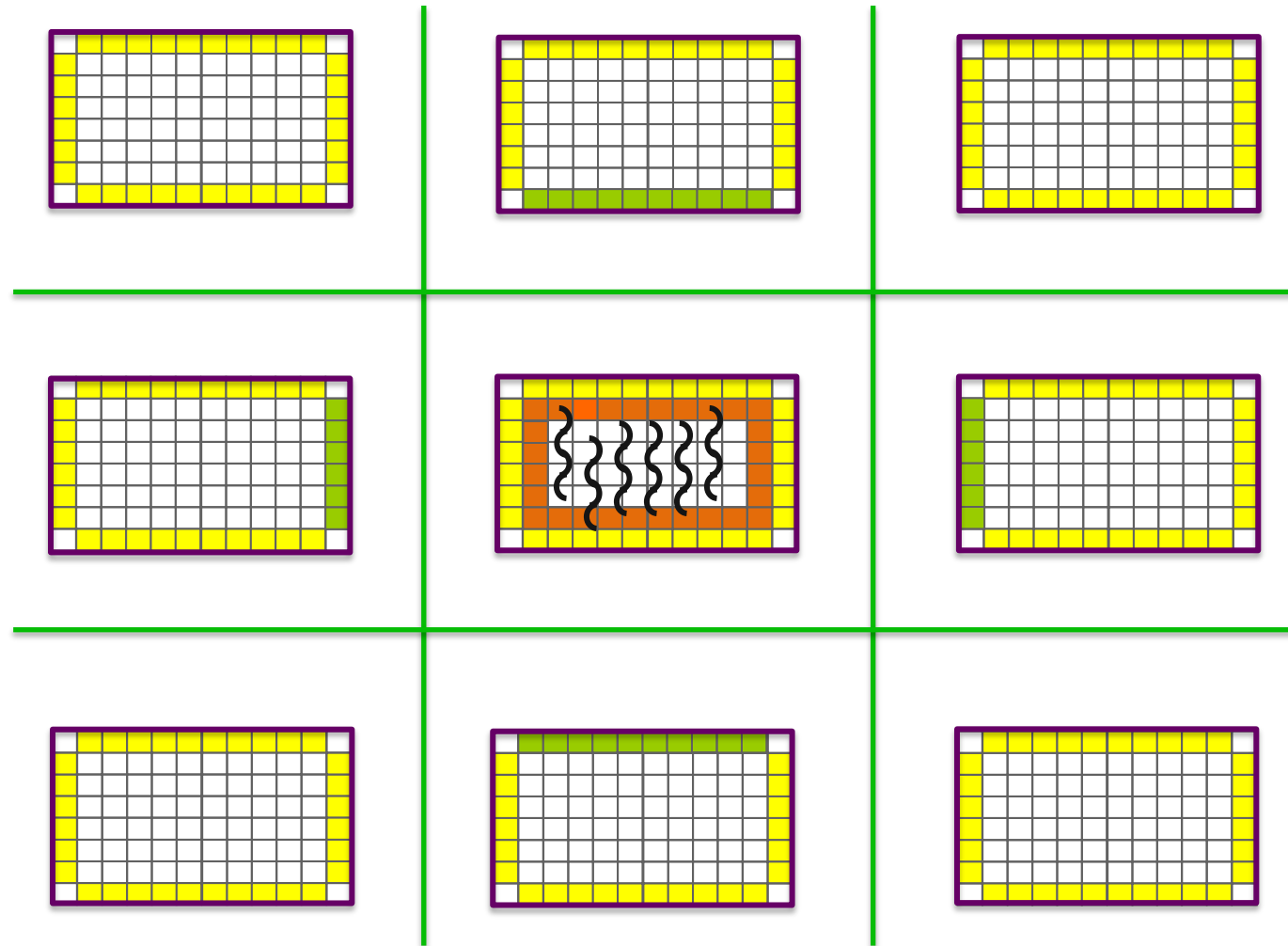
    #pragma omp parallel for
    for (i = 0; i < 100; i++) {
        target = rand();
        MPI_Win_lock(MPI_LOCK_EXCLUSIVE, target, 0, win);
        MPI_Put(..., win);
        MPI_Win_unlock(target, win);
    }

    /* Free MPI and RMA window */

    return 0;
}
```

Different threads can lock the same process causing multiple locks to the same target before the first lock is unlocked

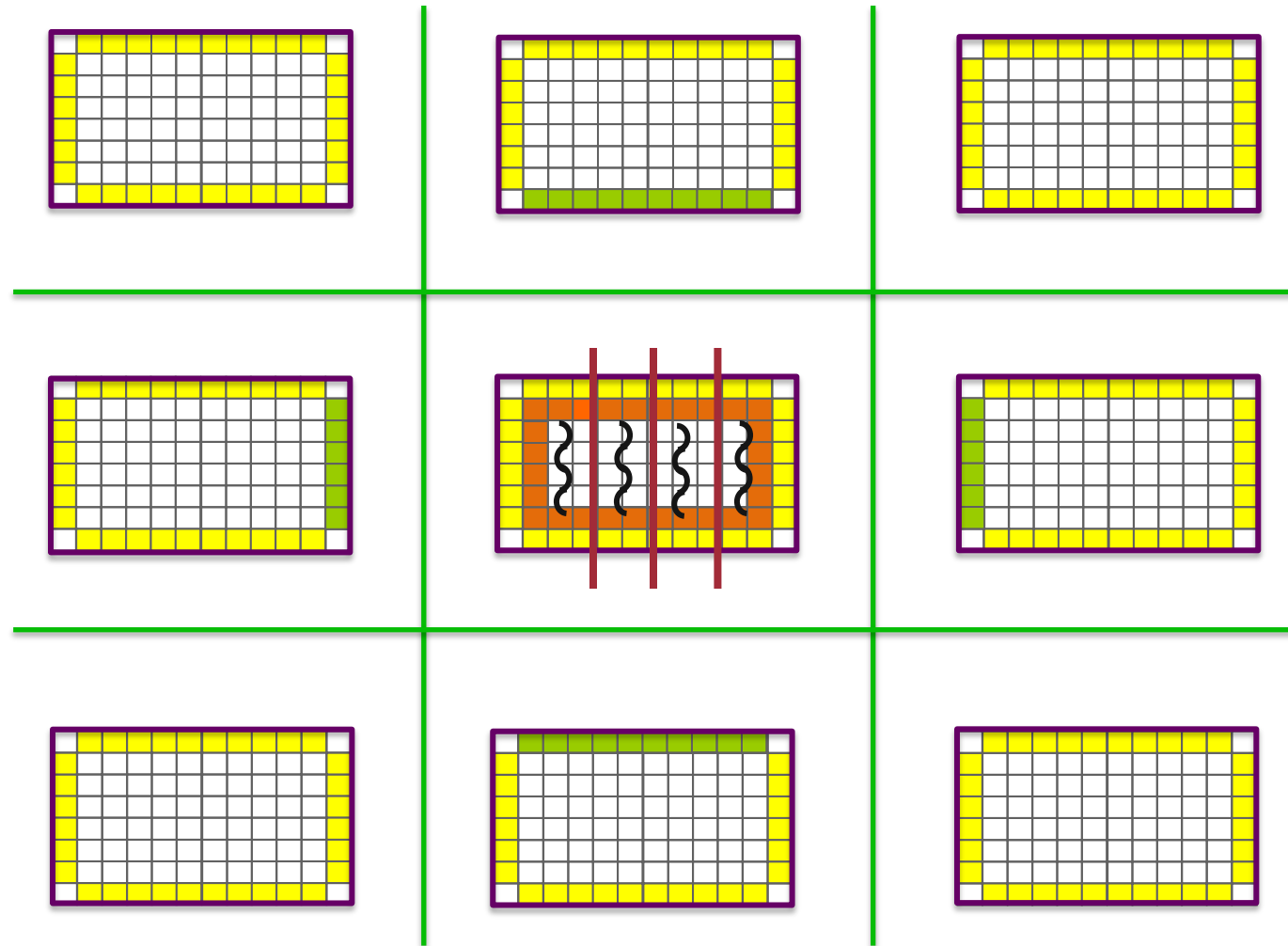
Exercise 1: Stencil in Funneled mode (1/2)



Exercise 1: Stencil in Funneled mode (2/2)

- Parallelize computation (OpenMP parallel for)
- Main thread does all communication
- *Start from derived_datatype/stencil.c*
- *Solution available in threads/stencil_funneled.c*

Exercise 2: Stencil in Multiple mode (1/2)



Exercise 2: Stencil in Multiple mode (2/2)

- Divide the process memory among OpenMP threads
- Each thread responsible for communication and computation
- *Start from threads/stencil_funneled.c*
- *Solution available in threads/stencil_multiple.c*

Exercise 3: BSPMM in Funneled mode

- Parallelize dgemm computation (OpenMP parallel for)
- Main thread does all communication
- *Start from rma/bspmm_counter.c*
- *Solution available in threads/bspmm_funneled.c*

Exercise 4: BSPMM in Multiple mode

- Each thread queries the next available block multiplication
- *Start from threads/bspmm_funneled.c*
- *Solution available in threads/bspmm_multiple.c*

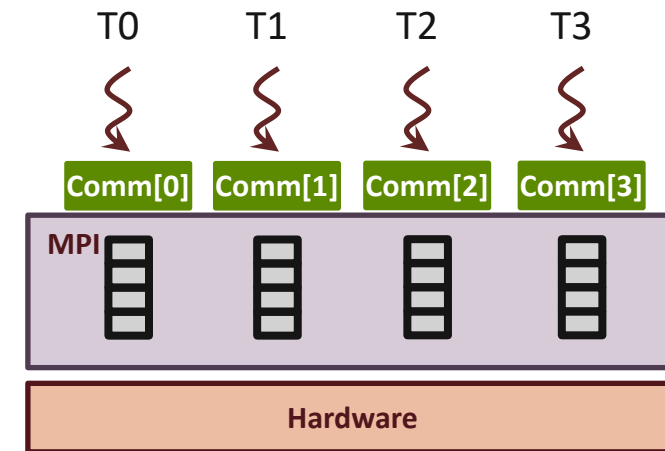
MPI+threads performance recommendations



Recommendation: Maximize independence between threads with communicators

- Each thread accesses to a **different communicator**
 - Each communicator may be associated with isolated resource in an MPI implementation

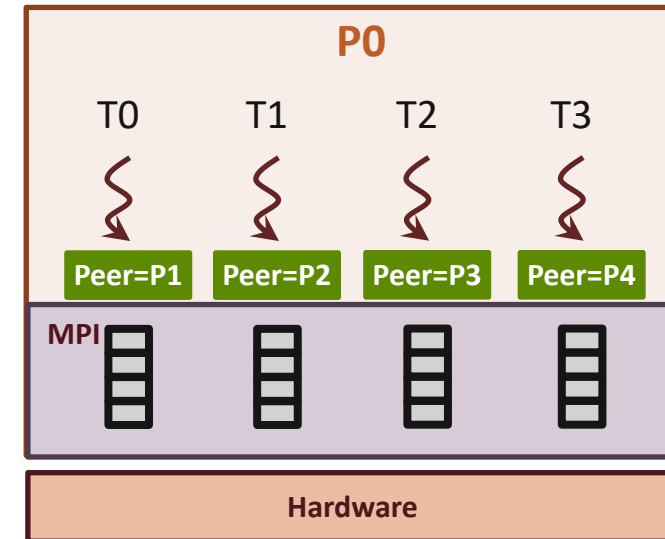
```
MPI_Comm *comms;  
int nthreads = omp_get_num_threads();  
comms = malloc(sizeof(MPI_Comm) * nthreads);  
  
for (i = 0; i < nthreads; i++)  
    MPI_Comm_dup(MPI_COMM_WORLD, &comms[i]);  
  
#pragma omp parallel  
{  
    int tid = omp_get_thread_num();  
    #pragma omp taskloop  
    for (i = 0; i < 100; i++)  
        MPI_Isend(..., comm[tid], &req[i]);  
}  
MPI_Waitall(100, req, ..);
```



Recommendation: Maximize independence between threads with ranks or tags (1/2)

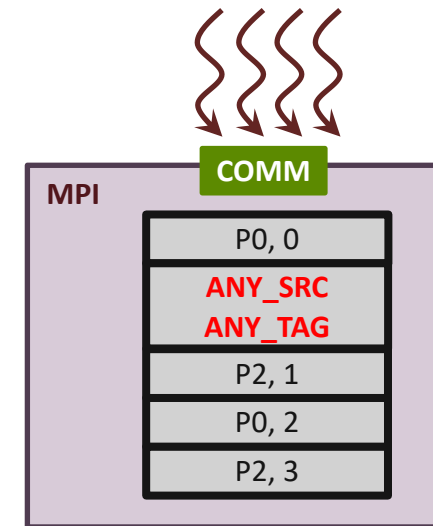
- Each thread communicates with **different peer_rank or tag**
 - MPI may assign isolated resource for different set of [peer_rank + tag]

```
#pragma omp parallel
{
    int tid = omp_get_thread_num();
    #pragma omp taskloop
    for (i = 0; i < 100; i++)
        MPI_Isend(..., peer_ranks[tid], tid,
                  comm, &req[i]);
}
MPI_Waitall(100, req, ..);
```



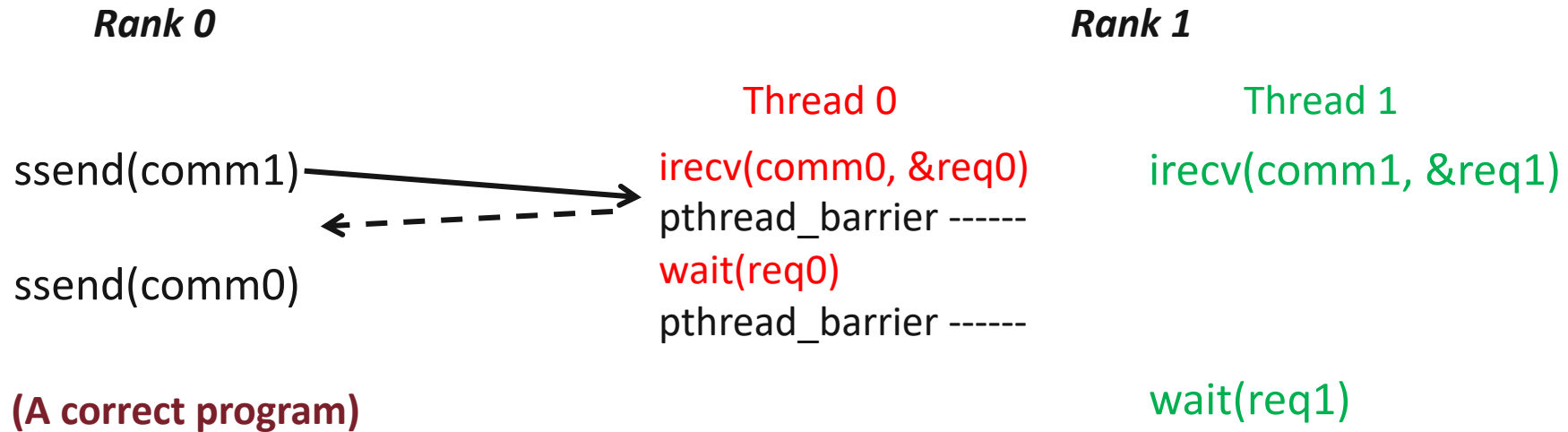
Recommendation: Maximize independence between threads with ranks or tags (2/2)

- Threads have to match all receive messages in sequential (e.g., a single receive-queue) if a **wildcard receive** may be posted
 - Ensure ordering of message matching
- Let MPI know if you do not use wildcard receive**
 - Info hints `mpi_assert_no_any_source`, `mpi_assert_no_any_tag` (already proposed to MPI standard)
 - MPI can get rid of the single receive-queue for the communicator



```
MPI_Info info;  
info = MPI_Info_create();  
MPI_Info_set(info,  
    "mpi_assert_no_any_source", "true");  
MPI_Comm_set_info(comm, info);  
MPI_Info_free(&info);  
/* Communicate without ANY_SOURCE */
```

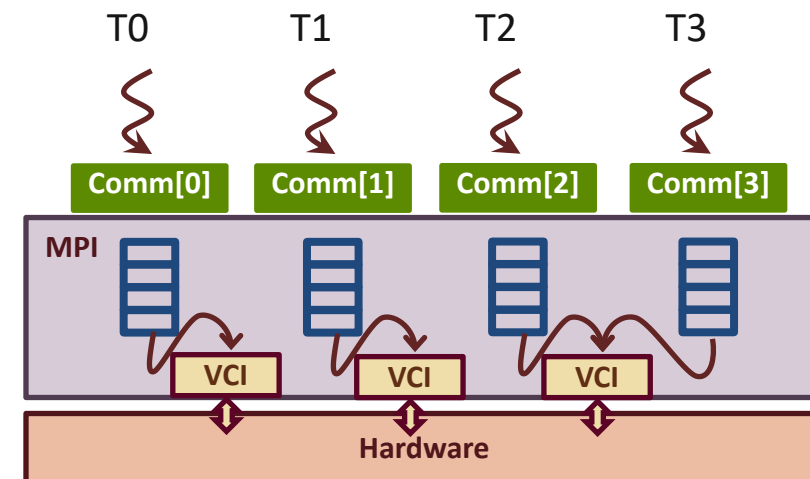
Communication Isolation Limitations



- **Progress:** A blocked thread will not prevent progress of other runnable threads on the same process
 - `ssend(comm1)` returns only after `irecv(comm1)` is posted
 - MPI may internally send handshake messages to synchronize
 - Thread 0 has to make progress for comm1 in `wait(req0)` (e.g., **access comm1's receive-queue**), to ensure `ssend(comm1)` can complete

Possible Optimizations MPI libraries can do

- Virtual Communication Interface (VCI)
 - Each VCI abstracts a set of network/shared-memory resources
 - Some networks support multiple VCIs: InfiniBand contexts, scalable endpoints over Intel Omni-Path
 - Traditional MPI implementation uses single VCI
 - Serializes all traffic
 - Does not fully exploit network hardware resources
- **Utilizing multiple VCIs to maximize independence** in communication
 - Separate VCIs per communicator or per RMA window
 - Distribute traffic between VCIs with respect to ranks, tags, and generally out-of-order communication
 - M-N mapping between Work-Queues and VCIs



Exercise 5: Stencil with Independent Communicators

- Divide the process memory among OpenMP threads
- Each thread responsible for communication and computation
- Each thread uses a different communicator
- *Start from threads/stencil_multiple.c*
- *Solution available in threads/stencil_multiple_ncomms.c*

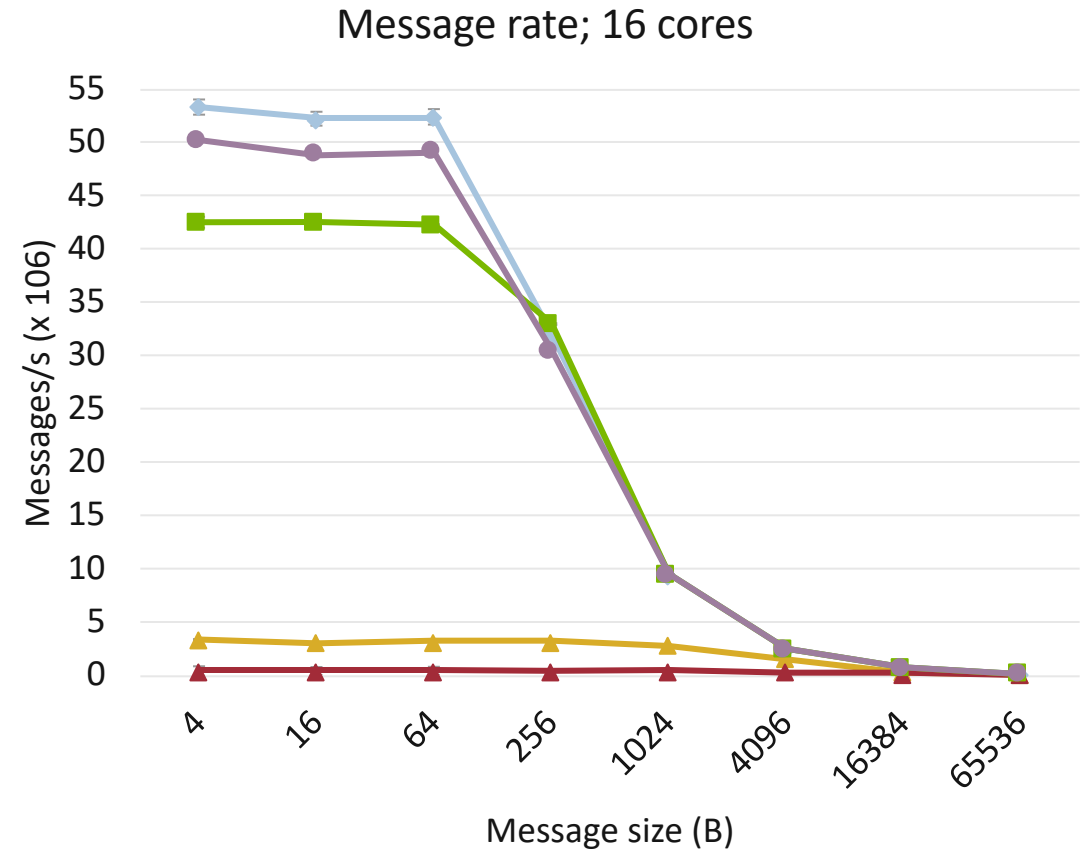
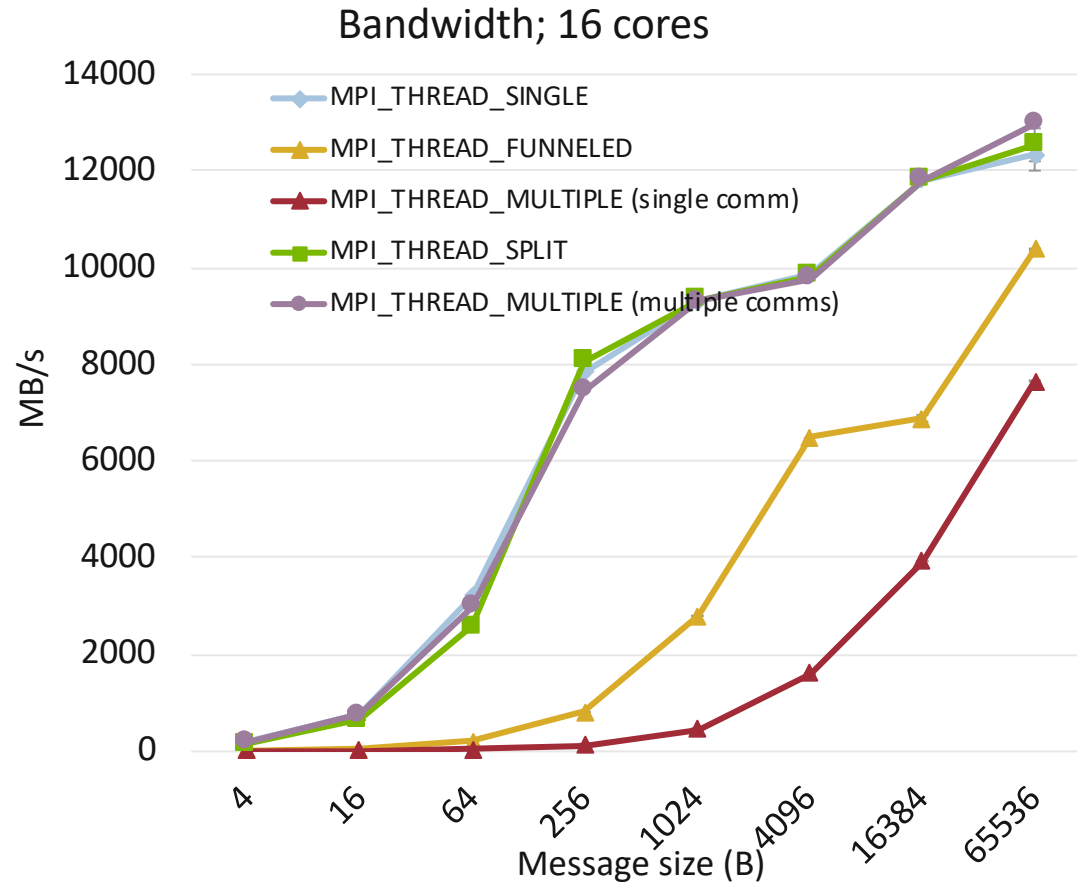
Exercise 6: BSPMM with overlapping windows

- Each thread queries the next available block multiplication
- Each thread uses a different window (overlapped) for accessing A and B global matrices
 - Have to use a single window for C matrix and global counter in order to ensure atomicity
- *Start from threads/bspmm_multiple.c*
- *Solution available in threads/bspmm_multiple_nwins.c*

MPI+threads optimizations in Intel MPI 2019

- Supported starting Intel MPI Library 2019 (for Linux).
- `MPI_Init_thread(MPI_THREAD_MULTIPLE)`
- Environment variables
 - `I_MPI_THREAD_SPLIT=1`
 - `I_MPI_THREAD_RUNTIME=openmp`
- `MPI_THREAD_SPLIT`
 - Only threads with the same thread id can communicate using distinct communicators, limiting space of possible communication patterns.
- Known issues
 - Using `MPI_PROC_NULL` with `I_MPI_THREAD_SPLIT=1` causes errors.
 - `MPI_Finalize` can take long with `I_MPI_THREAD_SPLIT=1`.

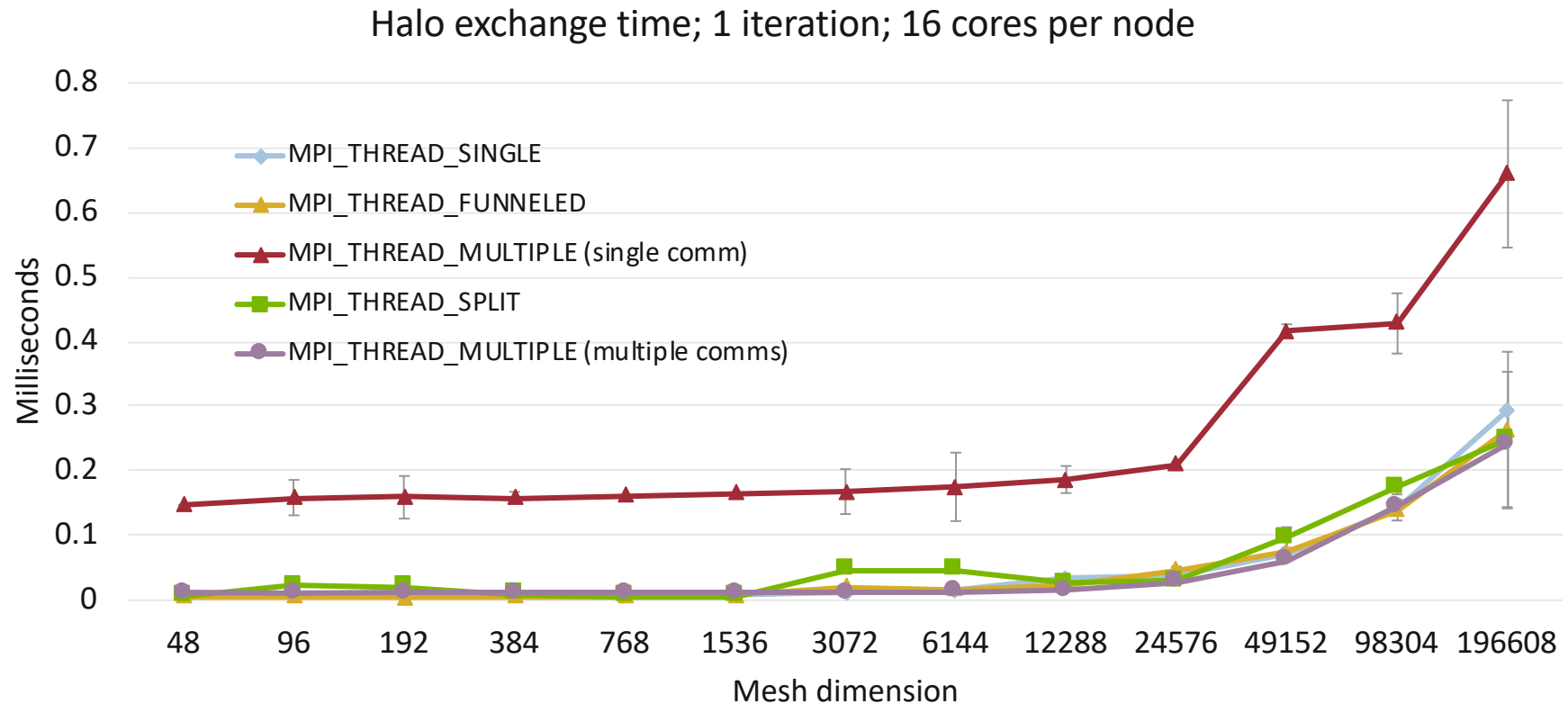
Exposing parallelism in micro-benchmarks



- All results, except “MPI_THREAD_MULTIPLE (multiple comms)” use Intel MPI 2019. The multiple comms version will be included into Intel MPI soon.

Exposing parallelism in stencil example

- Using 16 cores per node; 3 x 3 node grid.
- Halo exchange time of a rank on the central node.
- All results, except “MPI_THREAD_MULTIPLE (multiple comms)” use Intel MPI 2019. The multiple comms version will be included into Intel MPI soon.



Section Summary

- Hybrid MPI + “X” is a promising approach for large scale programming (e.g., MPI+Threads)
 - Less memory consumption
 - More efficient on-node data movement (load/store)
- MPI thread safety: SINGLE, FUNNELED, SERIALIZED, MULTIPLE
- Use `MPI_Init_thread` for threaded programs (i.e., not SINGLE)
- `THREAD_MULTIPLE` ordering & progress semantics
- Always maximize independence between threads in your program
 - Independent communicators, no wildcard, independent `peer_ranks` and tags

MPI Hybrid Programming: Accelerators

(Thanks to Jiri Kraus @ NVIDIA for several corrections and comments)



U.S. DEPARTMENT OF
ENERGY

Introduction

Accelerators are becoming increasingly popular in parallel computing

■ CPUs

- Task-sequential execution model (focus on latency)
- Small # of complex compute cores (out-of-order execution)
- Deep pipelines
- Large caches
- Branch prediction hardware

■ GPUs

- Data-parallel execution model (focus on throughput)
- Large # of simple compute elements (in-order execution)
- Small caches
- Shallow pipelines
- Large off-chip global High-Bandwidth Memory (HBM)
- High FLOPs/W and FLOPs/\$

Top500 Accelerators Based Systems (June 2020)

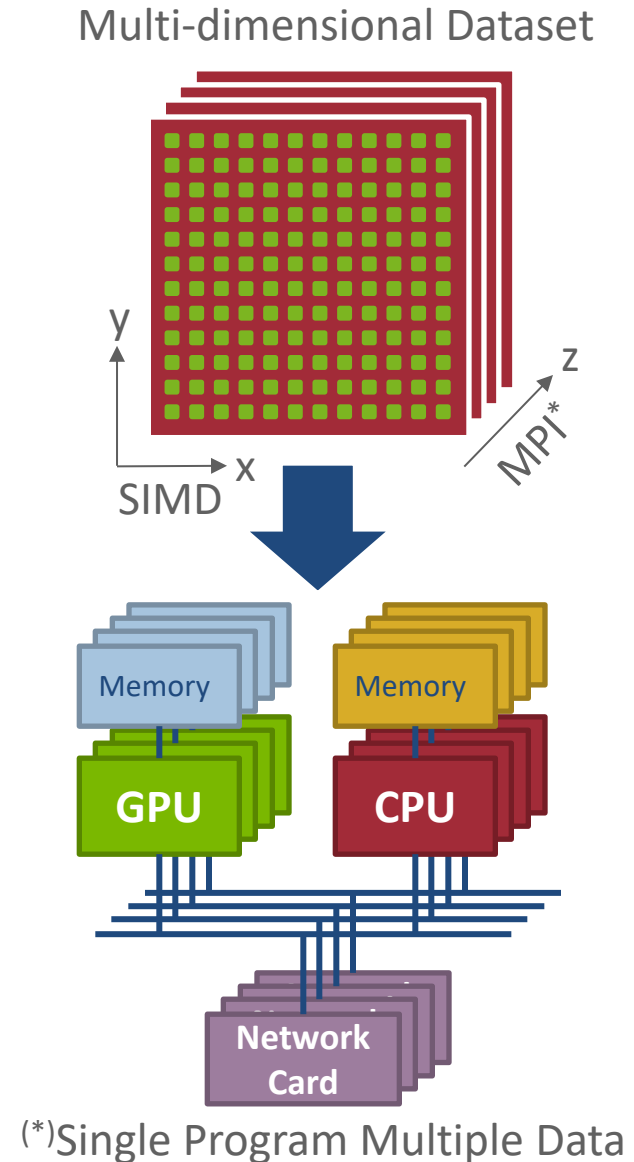
- #2 - Summit (ORNL USA)
 - NVIDIA Tesla V100
- #3 - Sierra (LLNL USA)
 - NVIDIA Tesla V100
- #6 - HPC5 (Eni S.p.A Italy)
 - NVIDIA Tesla V100
- #7 - Selene (NVIDIA Corporation USA)
 - NVIDIA Tesla A100
- #9 – Marconi-100 (CINECA Italy)
 - NVIDIA Tesla V100
- #10 - Piz Daint (CSCS Switzerland)
 - NVIDIA Tesla P100

Upcoming Exascale Accelerators Based Systems

- Aurora (ANL USA)
 - Intel based technology
 - <https://www.anl.gov/article/us-department-of-energy-and-intel-to-deliver-first-exascale-supercomputer>
- Frontier (ORNL USA)
 - AMD based GPU technology
 - <https://www.ornl.gov/news/us-department-energy-and-cray-deliver-record-setting-frontier-supercomputer-ornl>
- El Capitan (LLNL USA)
 - AMD based GPU technology
 - <https://www.llnl.gov/news/llnl-and-hpe-partner-amd-el-capitan-projected-worlds-fastest-supercomputer>

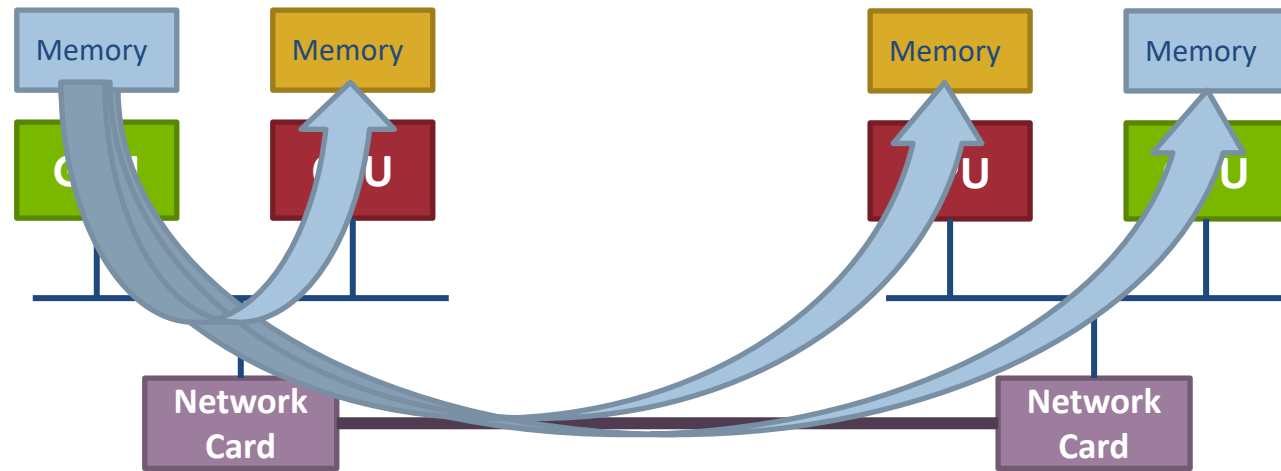
Programming Model for Accelerators

- **GPUs** are well suited for fine grain data level parallelism
- Shared Memory, Single Instruction Multiple Data (SIMD) model
- Many available compute platforms and programming frameworks (focus on their memory model and interaction with MPI)
 - NVIDIA CUDA (NVIDIA platform only)
 - AMD ROCm & HIP
 - OpenMP
 - OpenACC (mentioned but not covered)
 - OpenCL & SYCL



Interoperability with MPI

GPUs have separate physical memory subsystem
How to move data between GPUs with MPI?

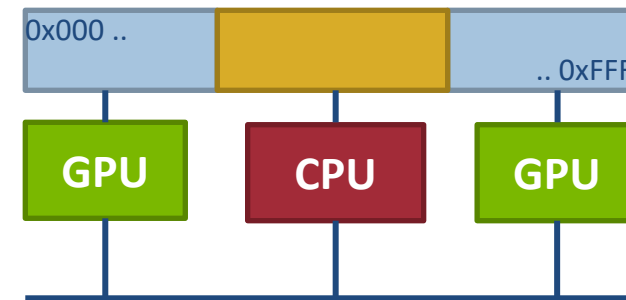


Real answer: It depends on what GPU library, what hardware and what MPI implementation you are using

Simple answer: For modern GPUs, “just like you would with a non-GPU machine”

Unified Virtual Addressing (UVA)

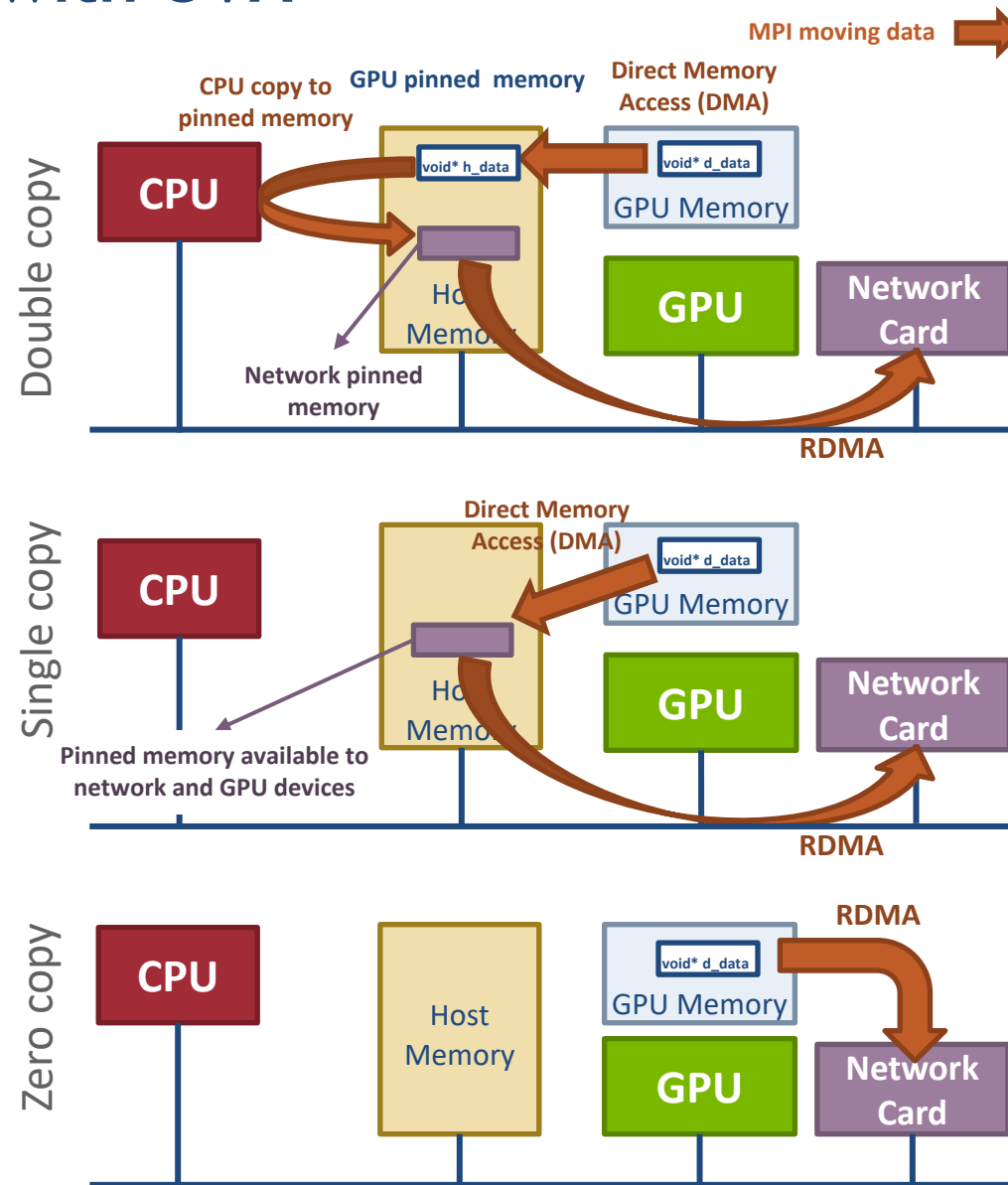
- UVA is a memory address management system supported in modern 64-bit architectures
 - Requires device driver support
- The same virtual address space is used for all processors, host or devices
- No distinction between host and device pointers
- The user can query the location of the data allocation given a pointer in the unified virtual address space and the appropriate GPU runtime library query APIs (“GPU-aware” MPI library)



UVA: Single virtual address space for the host and all devices

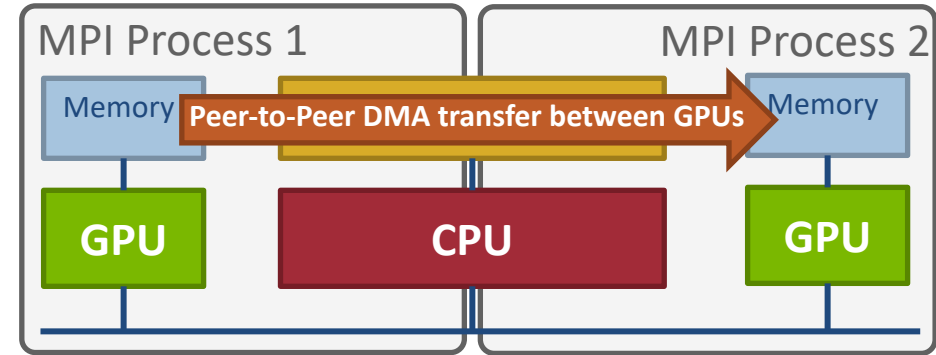
Remote Direct Memory Access with UVA

- Only GPU-enabled MPI implementations can take advantage of UVA
- User can pass device pointer to MPI
- MPI implementation can query for the owner (host or device) of the data
- If the data is on the device, **the MPI implementation can optimize data transfers**



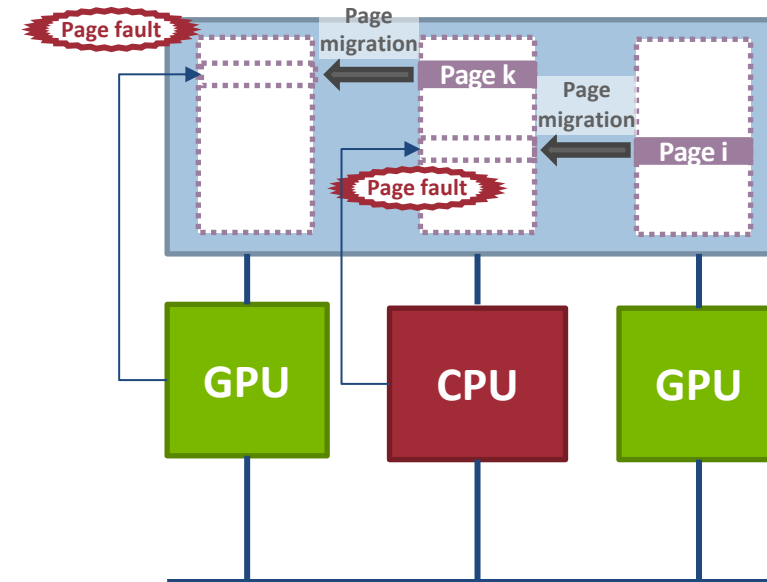
Intranode Communication with UVA

- Intranode Optimization
 - GPU peer-to-peer data transfers are possible
 - MPI can directly move data between GPU devices



Heterogeneous Memory Management (HMM)

- Next step towards the unification of heterogeneous memory spaces in the Linux Kernel (not yet available)
- Support started with version 4.14 through helper functions to be used by device drivers
 - Support paging in device for migrating memory between host and GPU
- Automatic data movement between host and GPU memories (called Unified Memory in CUDA)
 - Data is automatically migrated between host and GPU on page faults
 - Moving pages to GPU and back to host is similar to swap-out and swap-in of pages to and from disk



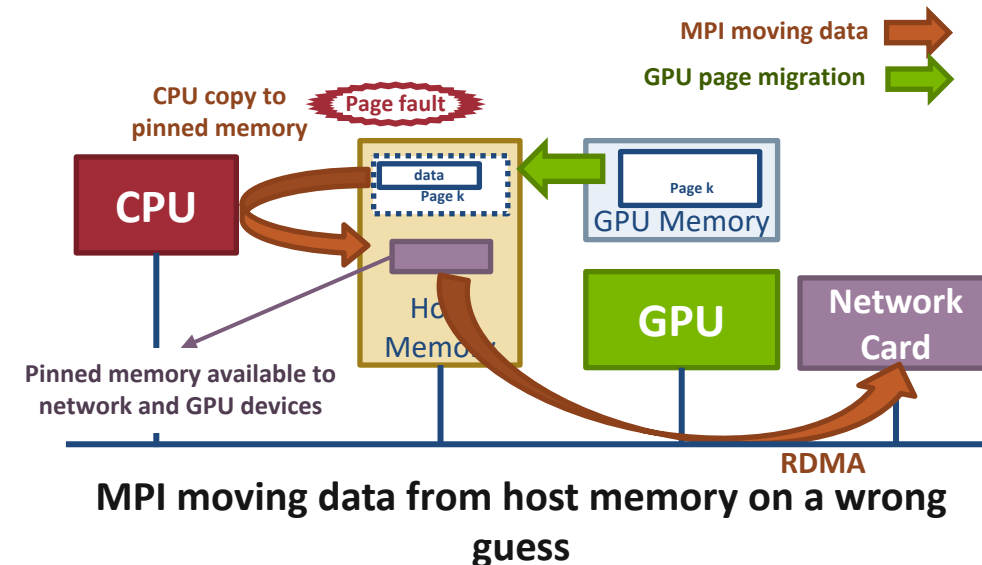
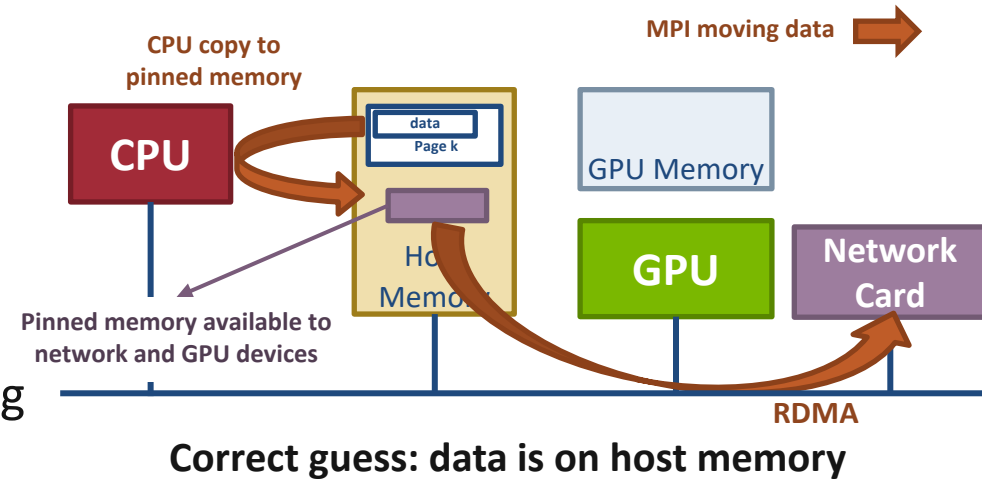
Single memory space accessible to all devices and host. Transparently managed heterogeneous memory.

MPI + HMM in a Nutshell

- In theory, any MPI implementation can transparently work with HMM
 - The MPI implementation can always assume that the data is on the host (or device)
 - GPUs take care of moving data between device and host memories
 - *Trying to register memory on the wrong device from the network should simply fail for HMM, but there have been reports of silent failures in this regard for CUDA, so you might need to be careful*
 - *Data in HMM can be corrupted if GPU updates pages during network transfer*
- In practice MPI implementations should never use HMM directly
 - Managed heterogeneous memory cannot be directly accessed by the network (**correctness issue**)
 - Virtual address cannot be pinned to a fixed physical memory region since GPU might need to migrate the data
 - Intermediate buffer is needed to copy data from HMM
- In any case MPI can never know in which device data is physically located
 - Data management is completely handled by GPU and can cause unnecessary data movement (**performance issue**)

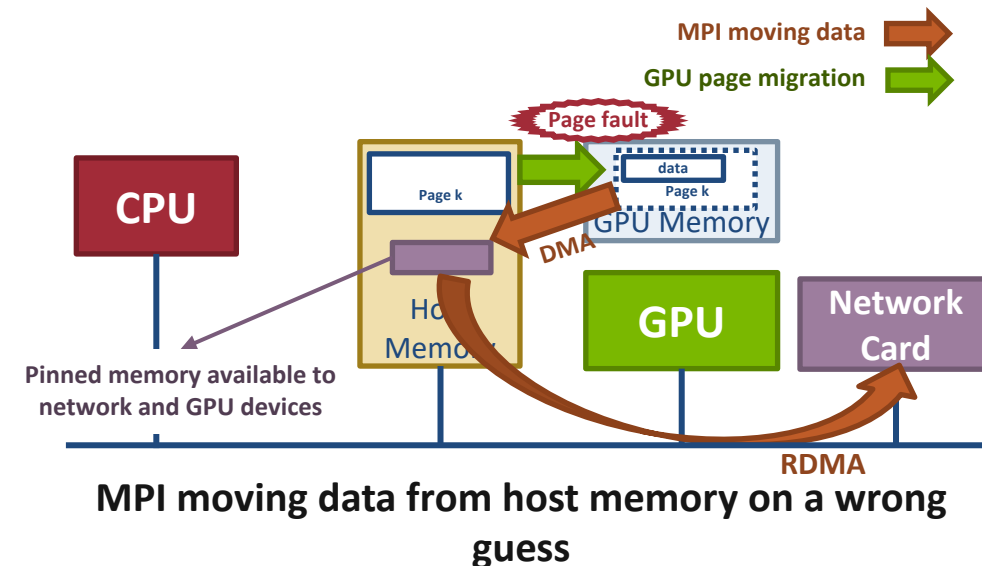
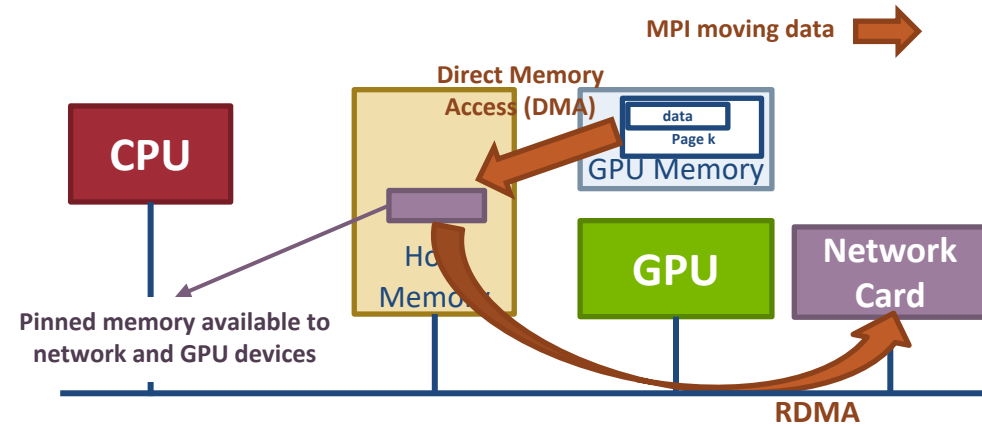
MPI + HMM Assuming Data on Host

- MPI can assume data is on host memory
- MPI copies data to network pinned memory
 - Network registration will fail
- On a correct guess
 - The copy will not trigger a page fault to bring data from GPU
- On incorrect guess
 - An **expensive page fault** will occur



MPI + HMM Assuming Data on GPU

- MPI can assume data is on some GPU device memory
- MPI would need to move data from the GPU device memory to network pinned memory
 - This can be either host or GPU memory (but not unified memory)
- On a correct guess
 - The copy will not trigger a page fault
- On incorrect guess
 - An **expensive page fault** will occur when accessing data on the GPU device memory
- Most MPI implementations assume memory to reside on the GPU



Compute Unified Device Architecture (CUDA)

(Thanks to CJ Newburn from NVIDIA for review and comments)



U.S. DEPARTMENT OF
ENERGY

Overview

- General-purpose parallel computing platform and programming model released by NVIDIA in 2006
- Provides a user library (*libcuda*), runtime (*libcudart*), device drivers and C/C++ compiler (NVCC)
- Programming language extensions for C/C++
 - Define C functions to run on the GPU (kernels) using the `__global__` declaration specifier
 - Kernels can be launched with different number of threads using the `<<<...>>>` execution syntax
 - Each thread executing the kernel is given a unique thread ID accessible from inside the kernel using the built-in `threadIdx` variable
- Support other languages such as FORTRAN

```
/* Kernel definition */
__global__ void gpu_kernel(double *in,
                           double *out)
{
    /* get indices from thread id */
    int i = threadIdx.x;
    int j = threadIdx.y;

    /* each thread performs work */
    out[i][j] = f(in, i, j);
}

int main()
{
    double *in_h, *out_h, *in_d, *out_d;
    in_h = malloc(size);
    out_h = malloc(size);
    cudaMalloc(&in_d, size);
    cudaMalloc(&out_d, size);

    cudaMemcpy(in_d, in_h, size,
               cudaMemcpyHostToDevice);

    /* kernel invocation with N threads */
    gpu_kernel<<<2,N>>>(in_d, out_d);

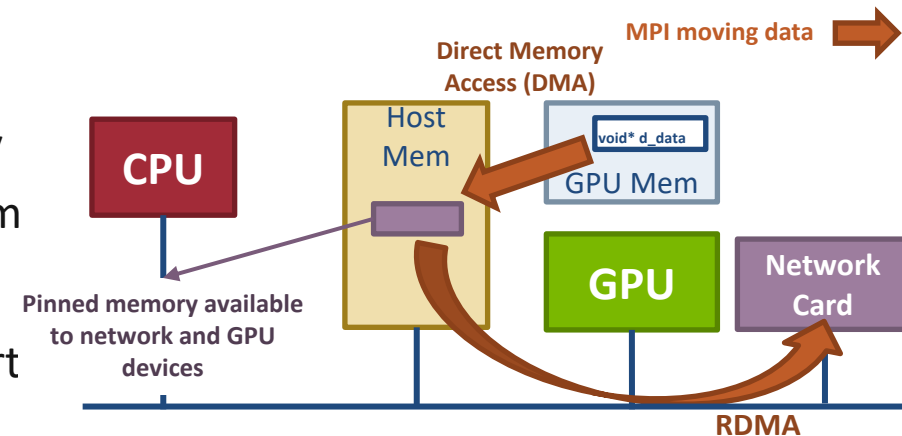
    cudaMemcpy(out_h, out_d, size,
               cudaMemcpyDeviceToHost);

    [...snip...]

    return 0;
}
```

MPI + GPUDirect (CUDA ≥ 4)

- GPUDirect 1.0 (Q2' 2010) allows pinned memory to be shared by GPU and NIC such that GPU can directly copy data in/out pinned memory and NIC can DMA data from it
- GPUDirect 2.0 (Peer-to-peer 2011) extends UVA support by allowing direct memory transfers between GPUs in the same node bypassing host completely



```
double *dev_buf;
cudaMalloc(&dev_buf, size);

if(my_rank == sender) {
    gpu_kernel<<<..>>>(dev_buf);
    MPI_Isend(dev_buf, size, MPI_DOUBLE, receiver, 0, comm, req);
} else {
    MPI_Recv(dev_buf, size, MPI_DOUBLE, sender, 0, comm, &status);
    gpu_kernel<<<..>>>(dev_buf);
}
```

MPI + GPUDirect RDMA (CUDA \geq 5)

- Technology introduced in 2013 with Kepler-class GPUs and CUDA-5
- GPU memory is directly accessible to third-party devices, including network interfaces (NIC driver talks to CUDA driver to register GPU memory)
- RDMA operations to/from the device memory are possible and completely bypass the host memory (zero copy)

```
double *dev_buf;  
cudaMalloc(&dev_buf, size);  
  
if(my_rank == sender) {  
    gpu_kernel<<<..>>>(dev_buf);  
    MPI_Isend(dev_buf, size, MPI_DOUBLE, receiver, 0, comm, req);  
} else {  
    MPI_Recv(dev_buf, size, MPI_DOUBLE, sender, 0, comm, &status);  
    gpu_kernel<<<..>>>(dev_buf);  
}
```

- Register memory to ensure CUDA memory operations complete before granting access to NIC:
 - cuPointerSetAttribute(..., CU_POINTER_ATTRIBUTE_SYNC_MEMOPS, ...)

MPI + Unified Memory (CUDA \geq 6)

- Unified Memory (UM) between host and device
 - CUDA kernel driver has a copy of the page table of the host and can handle pagefaults by migrating pages from host to device & vice versa
- UM-aware MPI implementations stage data into tmp buffer
- As mentioned before, performance can be bad
 - MPI never knows if pages are on host or device (can only guess)

```
double *buf;
cudaMallocManaged(&buf, size);

/* initialize buf in host ... */

if(my_rank == sender) {
    gpu_kernel<<<..>>>(buf);
    MPI_Isend(buf, size, MPI_DOUBLE, receiver, 0, comm, req);
} else {
    MPI_Recv(buf, size, MPI_DOUBLE, sender, 0, comm, &status);
    gpu_kernel<<<..>>>(buf);
}
```

MPI + CUDA Optimizations Historical Summary

Period	CUDA version	Major Features	MPI Optimization Space	MPI Implementation Requirements
After 2011	≥ 4.0 < 5.0	<ul style="list-style-type: none">• GPUDirect 1.0: RDMA can use GPU pinned memory• GPUDirect 2.0: GPU peer-to-peer DMA possible	<ul style="list-style-type: none">• Use DMA and RDMA without extra memory copies to temporary buffers	GPU-aware MPI implementations
After 2012	≥ 5.0	<ul style="list-style-type: none">• GPUDirect RDMA: GPU memory is directly accessible to third-party devices	<ul style="list-style-type: none">• Completely bypass host memory through RDMA to/from GPU memory	GPU-aware MPI implementations
After 2014	≥ 6.0	<ul style="list-style-type: none">• Unified Memory: shared memory between host and devices and automatic page migration	<ul style="list-style-type: none">• The hardware takes care of moving data between host and device memories• MPI optimizations are limited and need user hints	GPU-aware MPI implementations needed for performance

MPI + GPUDirect RDMA (Supported HW & SW)

- Mellanox Host Channel Adapters (HCA):
 - ConnectX-3, ConnectX-3 Pro, Connect IB, ConnectX-4, ConnectX-5
 - MLNX_OFED v2.1-x.x.x or later
 - Plugin Module to enable GPUDirect RDMA
- NVIDIA GPUs:
 - NVIDIA Tesla, Quadro K-Series or Tesla/Quadro P-Series
 - NVIDIA Driver
 - NVIDIA Runtime and Toolkit

Summary

- Accelerators are becoming increasingly important in HPC
- MPI is playing its role in enabling the usage of accelerators across distributed memory nodes
- The situation with MPI + GPU support is improving in both MPI implementations and in GPU hardware/software
 - For CUDA and ROCm P2P/RDMA support from GPU memory is enabled for contiguous datatypes through the UCX driver
 - For OpenCL/SYCL SVM can potentially enable similar optimizations as CUDA/ROCm but at the current state no such support is available and explicit data movement between host and device memory is still required
 - For OpenMP/OpenACC data movement is managed through directives but compilers can decide to leverage HMM

Exercise: Stencil with GPU

- Use GPU for updating the grid
 - Compute the grid on GPU
 - Pack/Unpack Halo regions to host memory for communication
- *Start from `accelerators/stencil_cuda_copy.cu`*
- *Solution available in `accelerators/stencil_cuda_direct.cu`*
- Special Environment Setup

For `stencil_cuda_copy.c`
+mvapich2
+cuda-10.0

For `stencil_cuda_direct.c`
+mvapich2-2.3.4-cuda10.0
+cuda-10.0

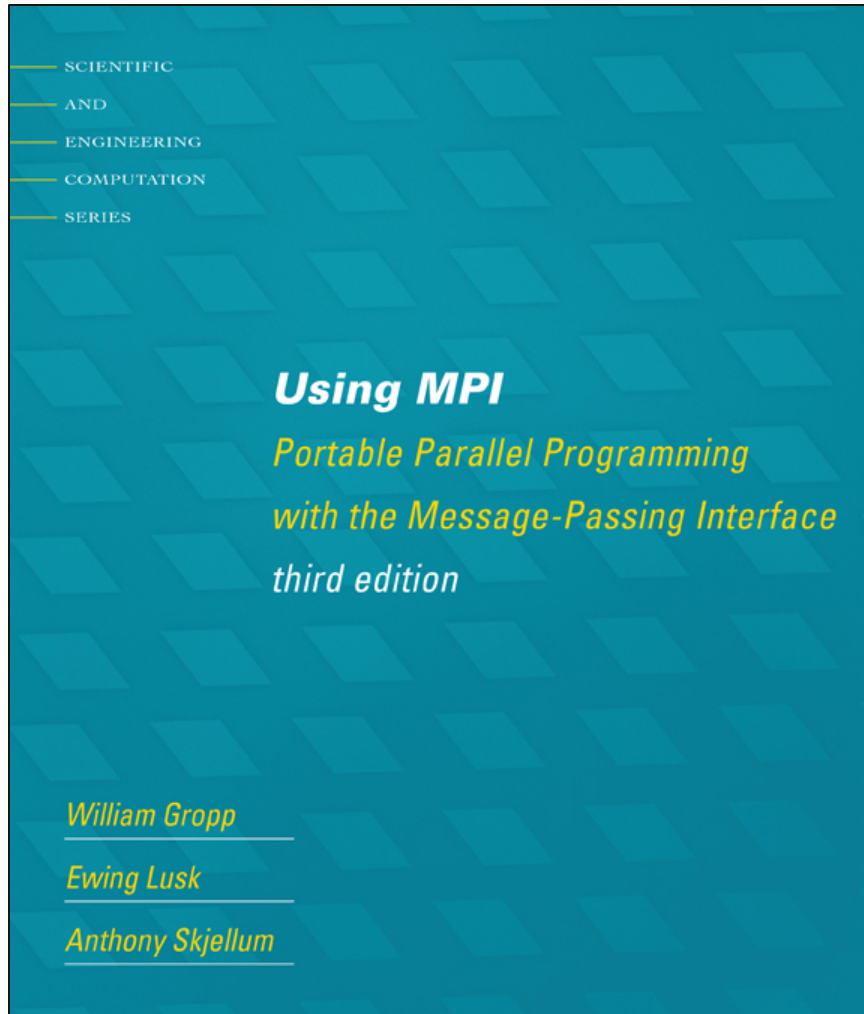
Concluding Remarks

- Parallelism is critical today, given that that is the only way to achieve performance improvement with the modern hardware
- MPI is an industry standard model for parallel programming
 - A large number of implementations of MPI exist (both commercial and public domain)
 - Virtually every system in the world supports MPI
- Gives user explicit control on data management
- Widely used by many scientific applications with great success
- Your application can be next!

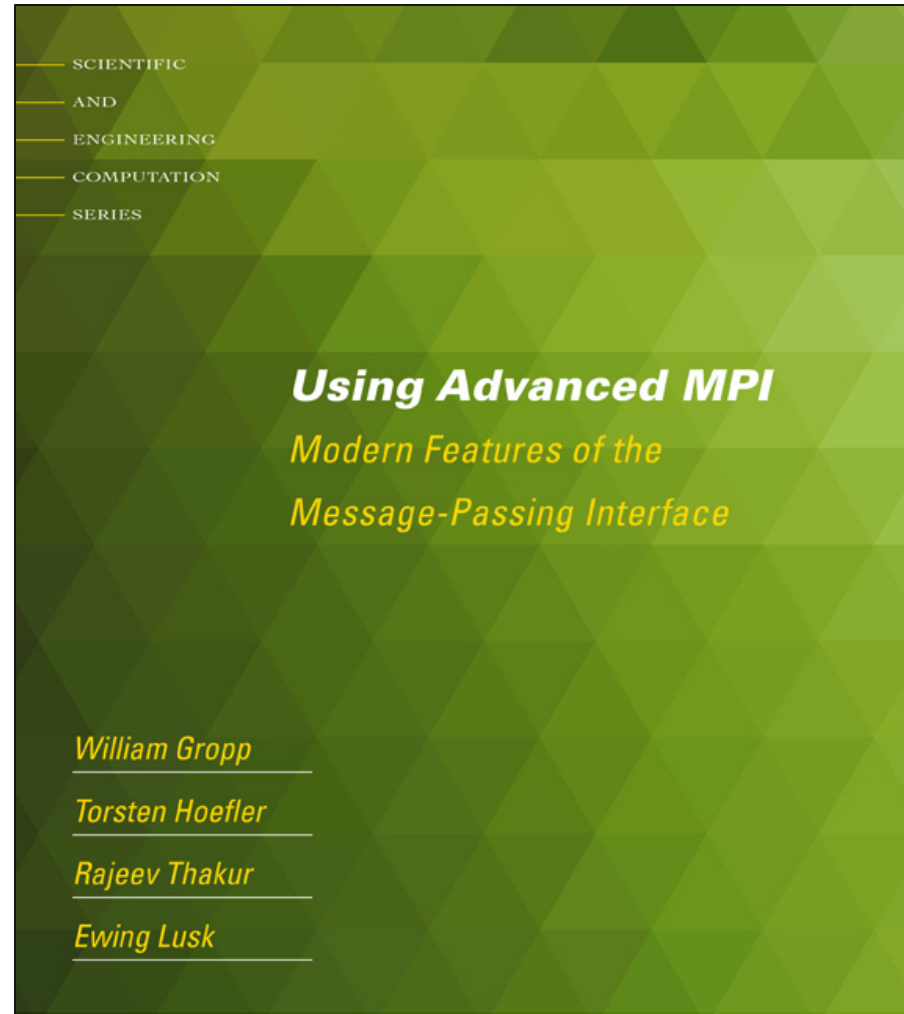
Web Pointers

- MPI standard : <http://www.mpi-forum.org/docs/docs.html>
- MPI Forum : <http://www.mpi-forum.org/>
- MPI implementations:
 - MPICH : <http://www.mpich.org>
 - MVAPICH : <http://mvapich.cse.ohio-state.edu/>
 - Intel MPI: <http://software.intel.com/en-us/intel-mpi-library/>
 - Microsoft MPI: <https://docs.microsoft.com/en-us/message-passing-interface/microsoft-mpi>
 - Open MPI : <http://www.open-mpi.org/>
 - IBM MPI, Cray MPI, HP MPI, TH MPI, ...
- Several MPI tutorials can be found on the web

Tutorial Books on MPI



Basic MPI

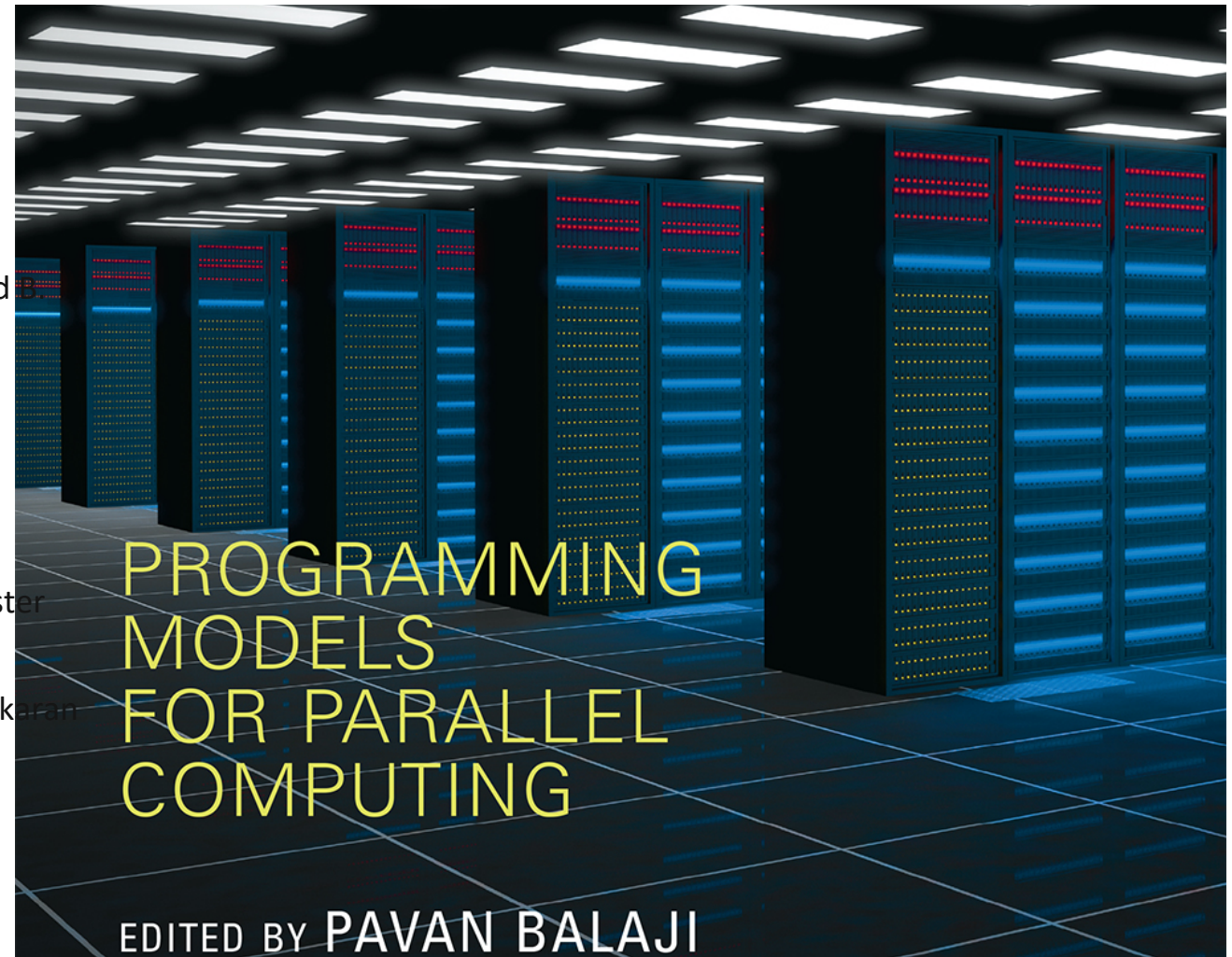


Advanced MPI, including MPI-3

Book on Parallel Programming Models

Edited by Pavan Balaji

- **MPI:** W. Gropp and R. Thakur
- **GASNet:** P. Hargrove
- **OpenSHMEM:** J. Kuehn and S. Poole
- **UPC:** K. Yelick and Y. Zheng
- **Global Arrays:** S. Krishnamoorthy, J. Daily, A. Vishnu, and Palmer
- **Chapel:** B. Chamberlain
- **Charm++:** L. Kale, N. Jain, and J. Lifflander
- **ADLB:** E. Lusk, R. Butler, and S. Pieper
- **Scioto:** J. Dinan
- **SWIFT:** T. Armstrong, J. M. Wozniak, M. Wilde, and I. Foster
- **CnC:** K. Knobe, M. Burke, and F. Schlimbach
- **OpenMP:** B. Chapman, D. Eachempati, and S. Chandrasekaran
- **Cilk Plus:** A. Robison and C. Leiserson
- **Intel TBB:** A. Kukanov
- **CUDA:** W. Hwu and D. Kirk
- **OpenCL:** T. Mattson



MPI for Scalable Computing

Tutorial at ATPESC, July 2020

Latest slides and code examples are available at

<https://anl.box.com/v/atpesc20-mpi-tutorial>

Pavan Balaji

Argonne National Laboratory

Email: balaji@anl.gov

Web: <http://www.mcs.anl.gov/~balaji>

William Gropp

University of Illinois, Urbana-Champaign

Email: wgropp@illinois.edu

Web: wgropp.cs.illinois.edu/

Yanfei Guo

Argonne National Laboratory

Email: yguo@anl.gov

Web: <https://www.mcs.anl.gov/~yguo/>

Rajeev Thakur

Argonne National Laboratory

Email: thakur@anl.gov

Web: <https://www.mcs.anl.gov/~thakur>



U.S. DEPARTMENT OF
ENERGY